Advanced Topics in Computer Vision

Dr. Lefei Zhang

zhanglefei@whu.edu.cn

April-2020

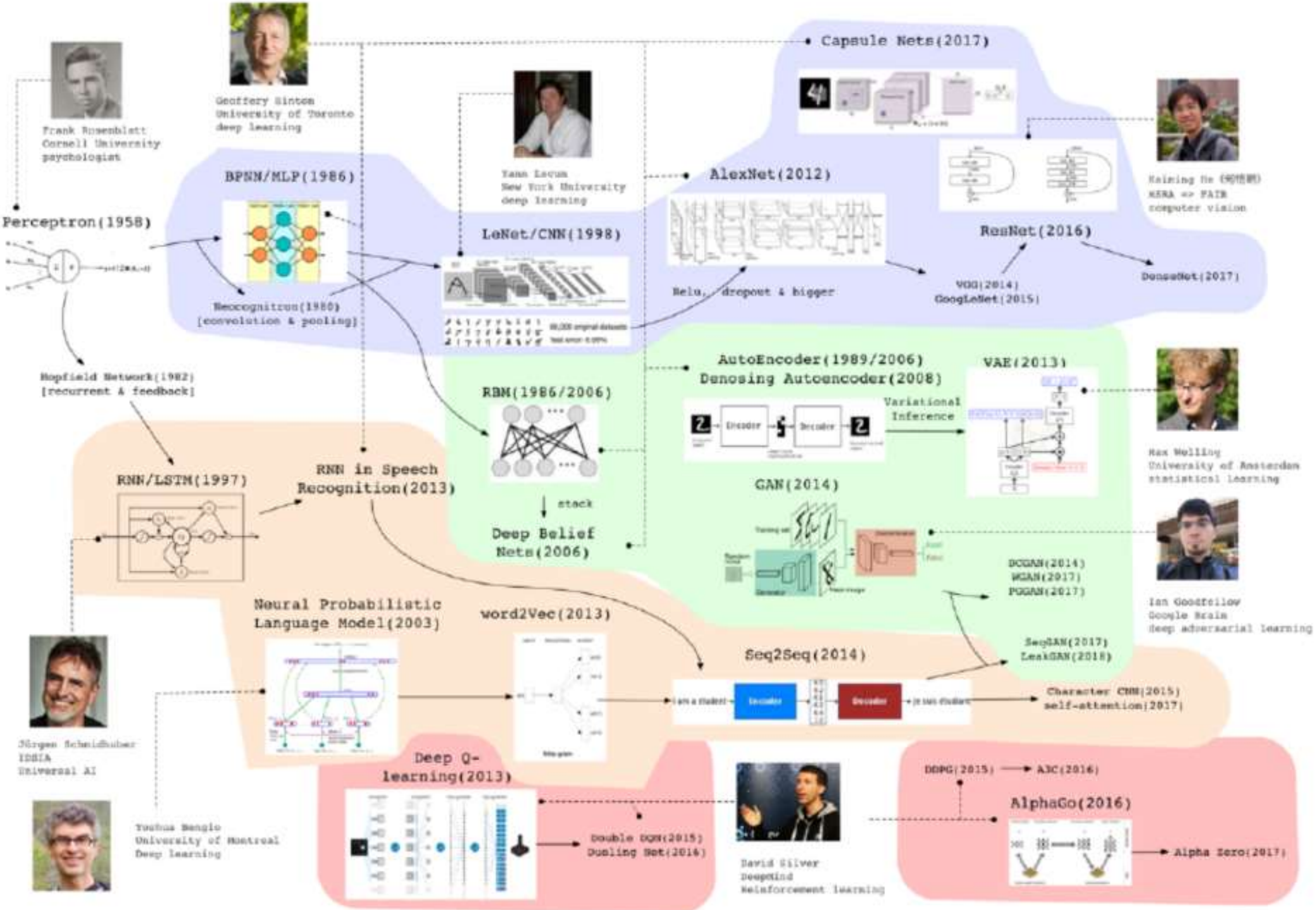
计算机视觉概念

计算机视觉（Computer Vision），顾名思义，是**分析、研究让计算机智能化，达到类似人类的双眼“看”的一门研究科学**。即对于客观存在的三维立体化的世界的理解，以及识别依靠智能化的计算机去实现。确切地说，计算机视觉技术就是利用了摄像机以及电脑替代人眼，使得计算机拥有人类的双眼，所具有的分割、分类、识别、跟踪、判别决策等功能。总之，计算机视觉系统就是创建了能够在2D的平面图像或者3D的三维立体图像的数据中，以获取所需要的“信息”的一个完整的人工智能系统。

计算机视觉发展历史

- 马尔计算视觉
- 多视几何与分层三维重建
- **基于学习的视觉（流形学习→深度学习）**

- CVPR
- ICCV
- ECCV



Frank Rosenblatt
Cornell University
psychologist

Geoffery Hinton
University of Toronto
deep learning

Yann Lecun
New York University
deep learning

Capsule Nets(2017)

Kaiming He (何恺明)
MSRA => FAIR
computer vision

Perceptron(1958)

BPNN/MLP(1986)

Neocognitron(1980)
[convolution & pooling]

LeNet/CNN(1998)

60,000 original datasets
Test error: 0.95%

AlexNet(2012)

Relu, dropout & bigger

ResNet(2016)

VGG(2014)
GoogLeNet(2015)

DenseNet(2017)

Hopfield Network(1982)
[recurrent & feedback]

RBM(1986/2006)

stack

Deep Belief
Nets(2006)

AutoEncoder(1989/2006)
Denosing Autoencoder(2008)

Encoder

Decoder

Variational
Inference

VAE(2013)

Max Welling
University of Amsterdam
statistical learning

GAN(2014)

DCGAN(2014)
WGAN(2017)
PGGAN(2017)

Ian Goodfellow
Google Brain
deep adversarial learning

RNN/LSTM(1997)

RNN in Speech
Recognition(2013)

Neural Probabilistic
Language Model(2003)

word2Vec(2013)

Seq2Seq(2014)

I am a student

Encoder

Decoder

je suis étudiant

SeqGAN(2017)
LeakGAN(2018)

Character CNN(2015)
self-attention(2017)

Jürgen Schmidhuber
IDSIA
Universal AI

Yoshua Bengio
University of Montreal
Deep learning

Deep Q-
learning(2013)

Double DQN(2015)
Dueling Net(2016)

David Silver
DeepMind
Reinforcement learning

DDPG(2015) → A3C(2016)

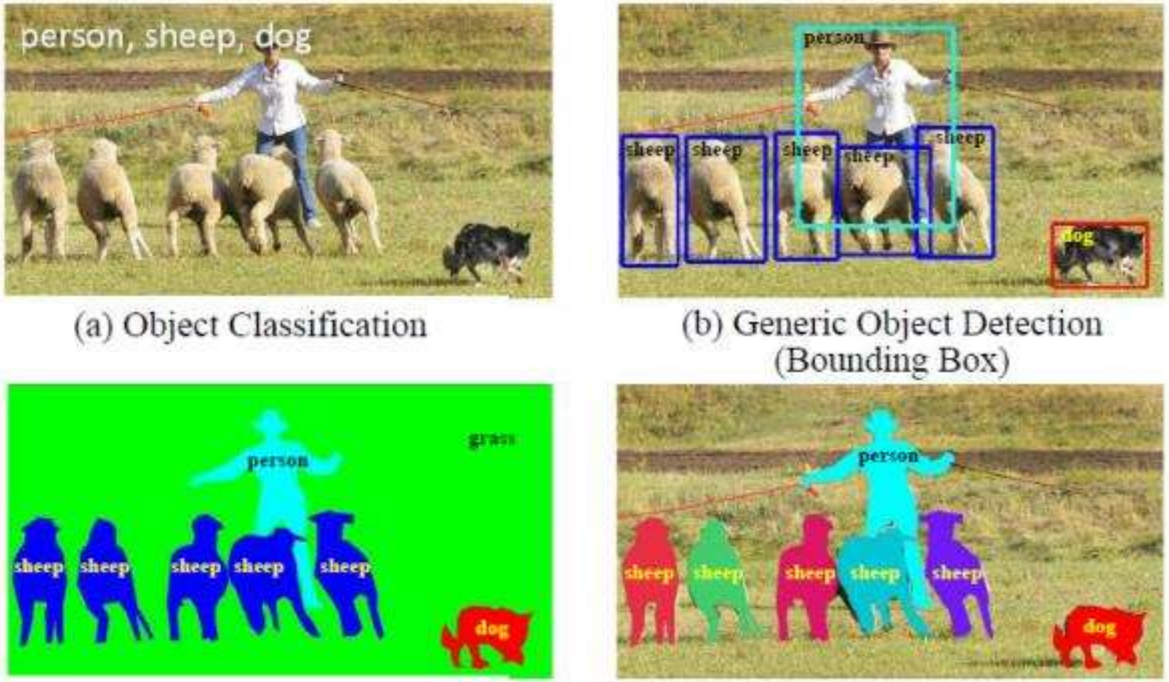
AlphaGo(2016)

Alpha Zero(2017)

1. 目标检测
2. 语义分割
3. 视频跟踪
4. ReID
5. 超分辨率
6. Inpainting
7. OCR
8. GAN

目标检测

目标检测任务的输入是一张图像，**输出是图像中的物体位置和类别**，如下图所示，位置可通过Bounding Box描述，也可描述为像素的集合。



(a) Object Classification

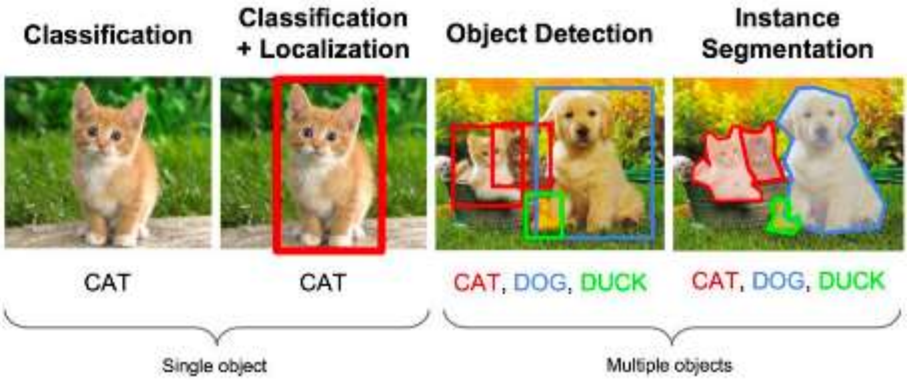
(b) Generic Object Detection (Bounding Box)

(c) Semantic Segmentation

(d) Object Instance Segmetation

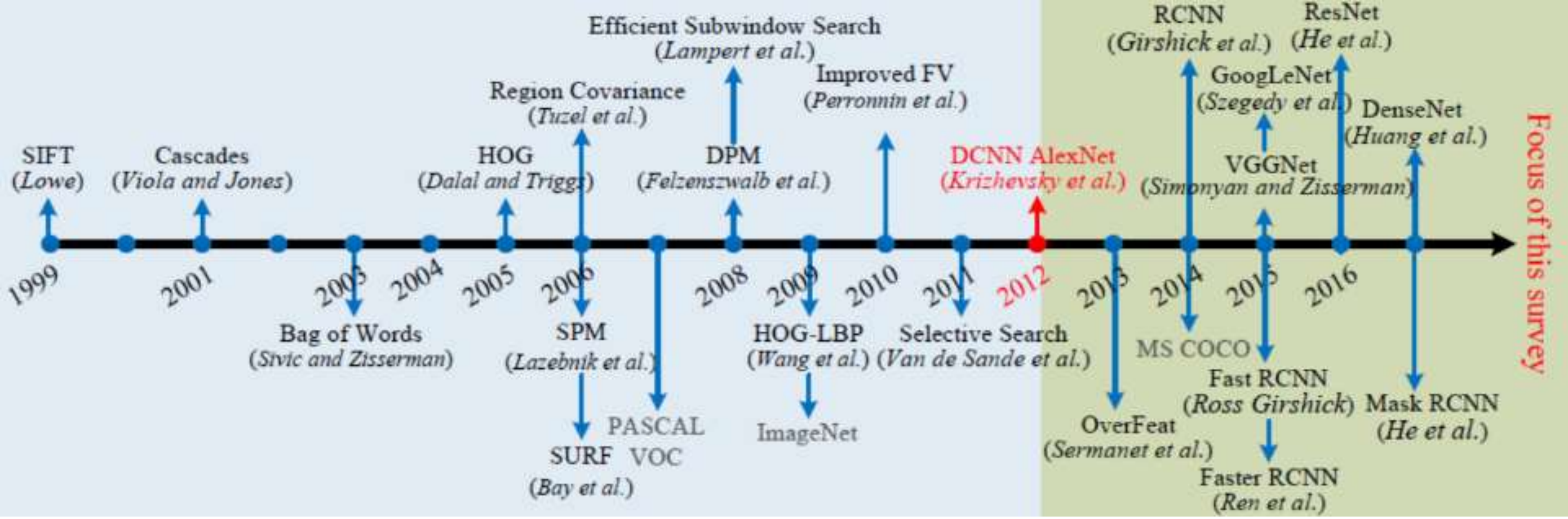
为了确定图片中物体的位置和类别，要面临很多挑战，一个好的检测器要做到**定位准确、分类准确、效率高**，需要对光照、形变、尺度、视角、尺寸、姿态、遮挡、模糊、噪声等情况鲁棒，需要能容忍可能存在的较大的类内差异，又能区分开较小的类间差异，同时还要保证高效。

目标检测

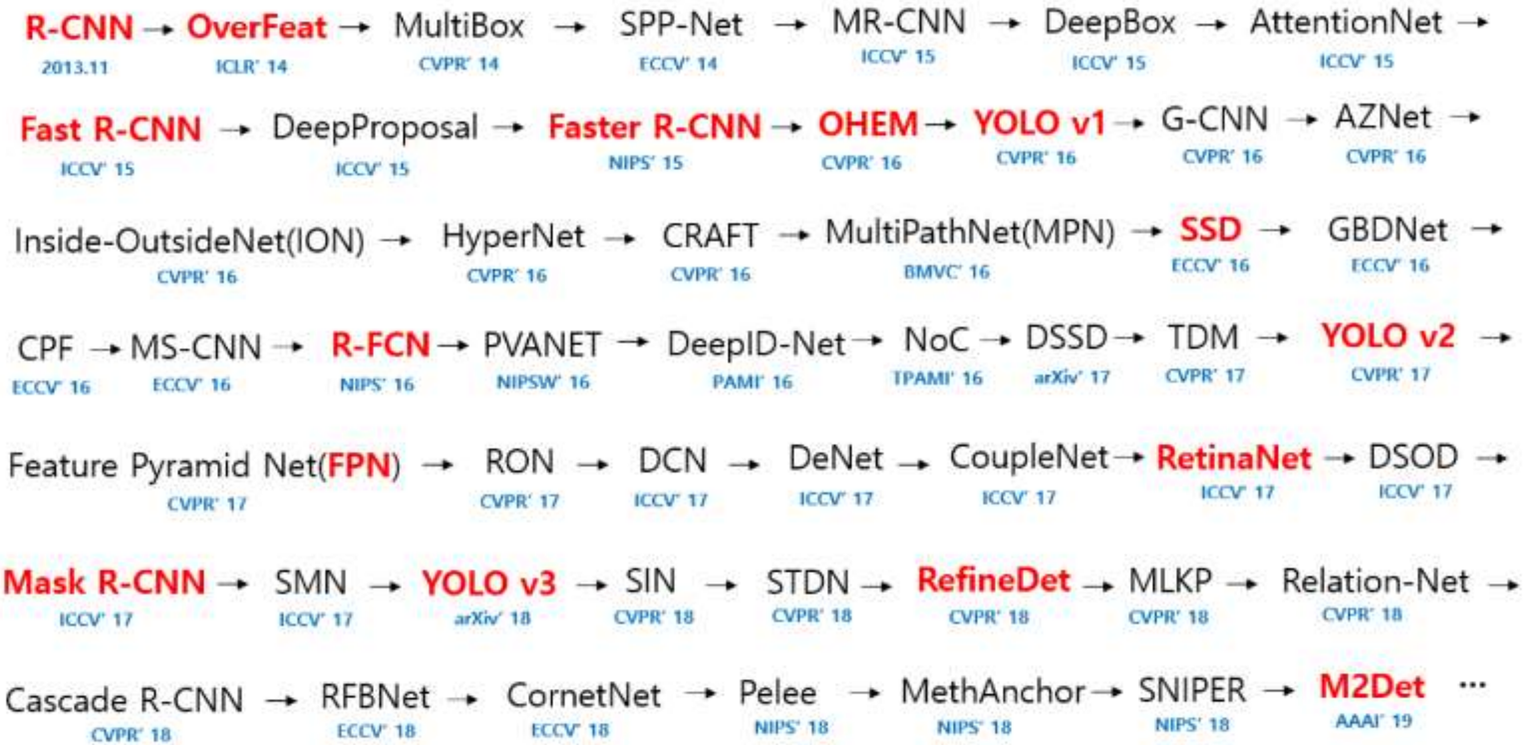


目标检测：定位图像中物体的位置，并在该物体周围绘制边界框

常用数据集	COCO、Pascal VOC、ImageNet等
常用指标	mAP，取各个类别的AP值的平均值得到mAP[Mean Average Precision]
	IoU，Intersection over Union，表示预测框与真实框之间的重叠程度
	FPS，每秒内可以处理的图片数量
代表算法	1.Ali Farhadi(华盛顿大学)——YOLO系列(You Only Look Once)
	2.Microsoft Research——R-CNN、Fast R-CNN、Faster R-CNN
	3.Facebook AI Research——Mask R-CNN、RetinaNet、FPN
	4.Google Inc.——SSD(Single Shot MultiBox Defender)、AmoebaNet
	5.XiaogangWang(香港中文大学)——CC-Net



SIFT (Lowe)
Cascades (Viola and Jones)
HOG (Dalal and Triggs)
Region Covariance (Tuzel et al.)
Efficient Subwindow Search (Lampert et al.)
DPM (Felzenszwalb et al.)
Improved FV (Perronnin et al.)
DCNN AlexNet (Krizhevsky et al.)
RCNN (Girshick et al.)
GoogLeNet (Szegedy et al.)
VGGNet (Simonyan and Zisserman)
ResNet (He et al.)
DenseNet (Huang et al.)
1999
2001
2003
2004
2005
2006
2008
2009
2010
2011
2012
2013
2014
2015
2016
Bag of Words (Sivic and Zisserman)
SPM (Lazebnik et al.)
SURF (Bay et al.)
PASCAL VOC
HOG-LBP (Wang et al.)
ImageNet
Selective Search (Van de Sande et al.)
OverFeat (Sermanet et al.)
MS COCO
Fast RCNN (Ross Girshick)
Faster RCNN (Ren et al.)
Mask RCNN (He et al.)
Focus of this survey



R-CNN (2013.11) → OverFeat (ICLR' 14) → MultiBox (CVPR' 14) → SPP-Net (ECCV' 14) → MR-CNN (ICCV' 15) → DeepBox (ICCV' 15) → AttentionNet (ICCV' 15) →

Fast R-CNN (ICCV' 15) → DeepProposal (ICCV' 15) → Faster R-CNN (NIPS' 15) → OHEM (CVPR' 16) → YOLO v1 (CVPR' 16) → G-CNN (CVPR' 16) → AZNet (CVPR' 16) →

Inside-OutsideNet(ION) (CVPR' 16) → HyperNet (CVPR' 16) → CRAFT (CVPR' 16) → MultiPathNet(MPN) (BMVC' 16) → SSD (ECCV' 16) → GBDNet (ECCV' 16) →

CPF (ECCV' 16) → MS-CNN (ECCV' 16) → R-FCN (NIPS' 16) → PVANET (NIPSW' 16) → DeepID-Net (PAMI' 16) → NoC (TPAMI' 16) → DSSD (arXiv' 17) → TDM (CVPR' 17) → YOLO v2 (CVPR' 17) →

Feature Pyramid Net(FPN) (CVPR' 17) → RON (CVPR' 17) → DCN (ICCV' 17) → DeNet (ICCV' 17) → CoupleNet (ICCV' 17) → RetinaNet (ICCV' 17) → DSOD (ICCV' 17) →

Mask R-CNN (ICCV' 17) → SMN (ICCV' 17) → YOLO v3 (arXiv' 18) → SIN (CVPR' 18) → STDN (CVPR' 18) → RefineDet (CVPR' 18) → MLKP (CVPR' 18) → Relation-Net (CVPR' 18) →

Cascade R-CNN (CVPR' 18) → RFBNet (ECCV' 18) → CornetNet (ECCV' 18) → Pelee (NIPS' 18) → MethAnchor (NIPS' 18) → SNIPER (NIPS' 18) → M2Det (AAAI' 19) ···

- **Two stage detection framework**：含Region Proposal，先获取ROI，然后对ROI进行识别和回归bounding box，以RCNN系列方法为代表。
- **One stage detection framework**：不含Region Proposal，将全图grid化，对每个grid进行识别和回归，以YOLO系列方法为代表。

- **主干网络 (Network Backbone) AlexNet→GoogLeNet→ResNet101→SENet**

- **多尺度 (Multiscale Object Detection)**

  1. Detecting with combined features of multiple CNN layers
  2. Detecting at multiple CNN layers;
  3. Combinations of the above two methods

- **目标几何形变**

  1. Deformable Part based Models (DPMs)
  2. Deformable Convolutional Networks (DCN)

- **上下文信息(Context Modeling)**

  1. Semantic context
  2. Spatial context
  3. Scale context

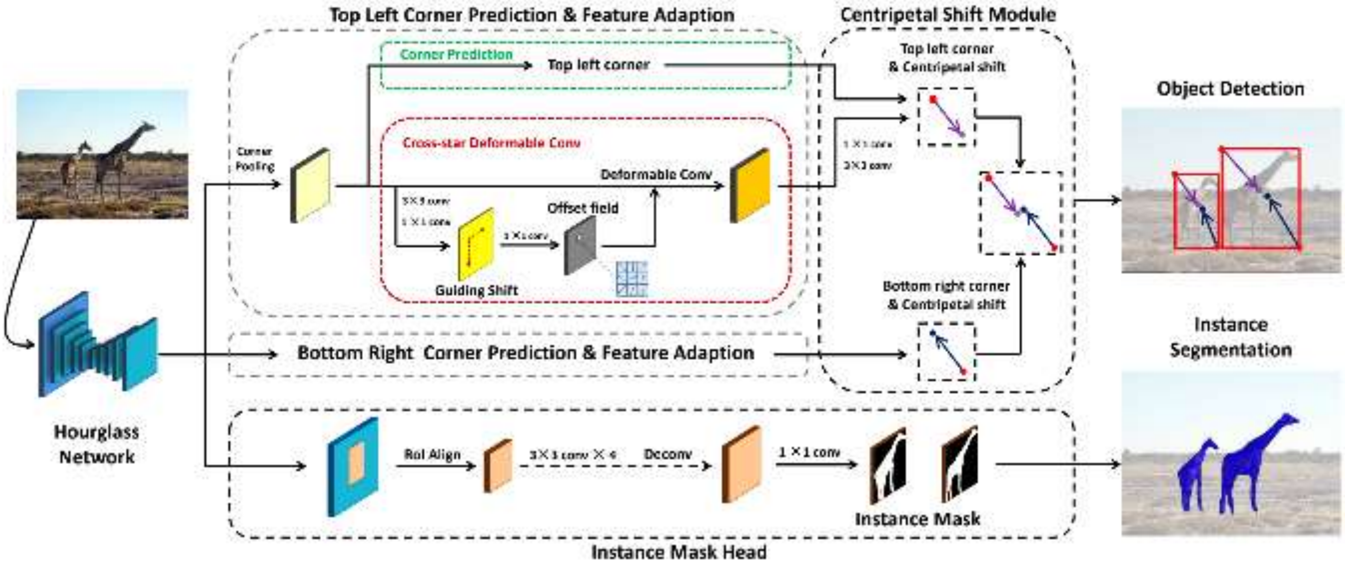
- **Detection Proposal Methods**

IJCV 2020. Deep Learning for Generic Object Detection: A Survey

- 目标检测 #1

CentripetalNet: Pursuing High-quality Keypoint Pairs for Object Detection

- Abstract & Overview



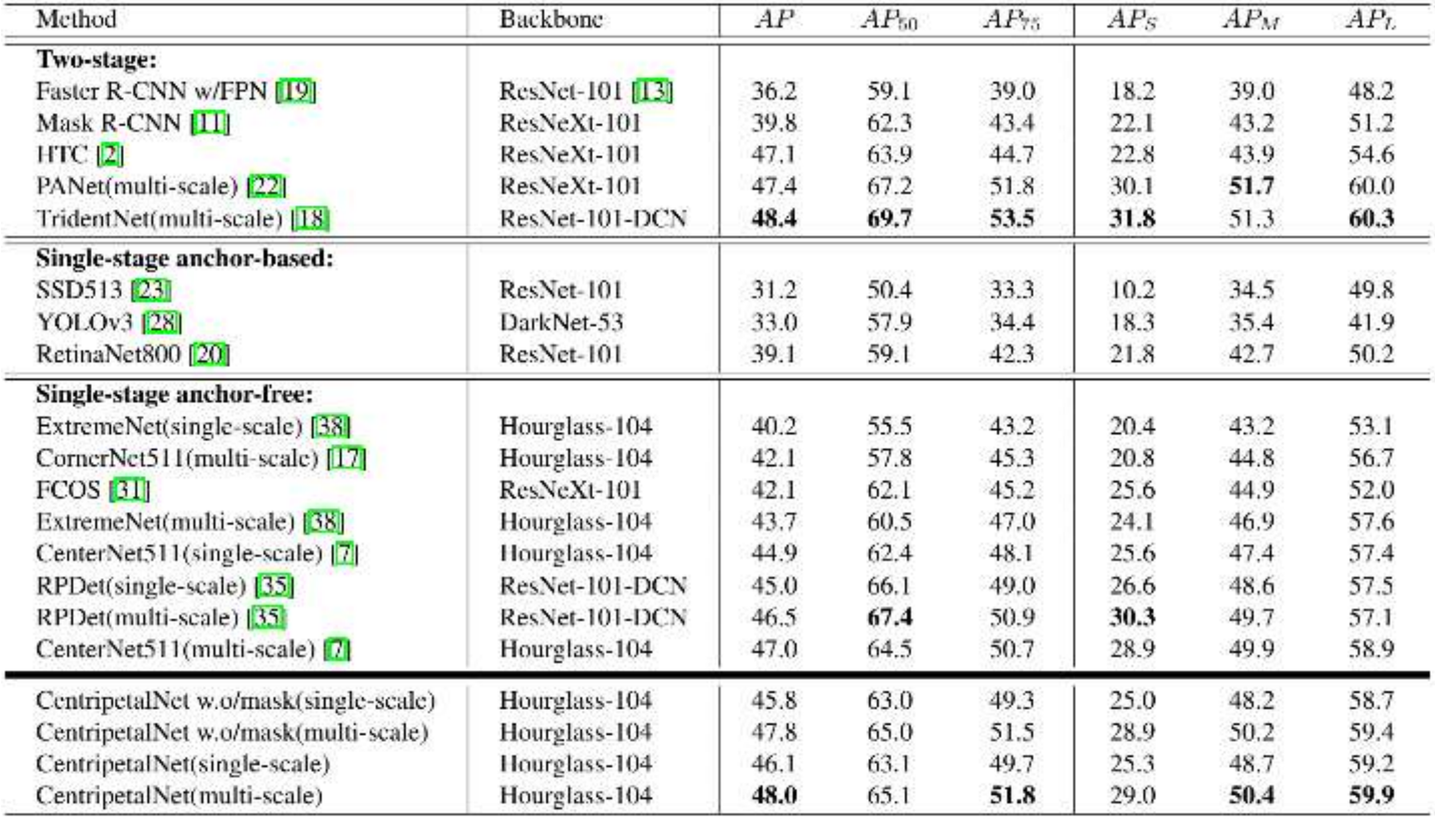
Keypoint-based detectors have achieved pretty-well performance. However, incorrect keypoint matching is still widespread and greatly affects the performance of the detector. In this paper, we propose **CentripetalNet** which uses centripetal shift to pair corner keypoints from the same instance. CentripetalNet predicts the position and the centripetal shift of the corner points and matches corners whose shifted results are aligned. Combining position information, our approach matches corner points more accurately than the conventional embedding approaches do. Corner pooling extracts information inside the bounding boxes onto the border. To make this information more aware at the corners, we design a **cross-star deformable convolution network** to conduct feature adaption. Furthermore, we explore instance segmentation on anchor-free detectors by equipping our CentripetalNet with a mask prediction module. On MS-COCO test-dev, our CentripetalNet not only outperforms all existing anchor-free detectors with an AP of 48.0%.

CVPR 2020. Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang*, Pengju Ren, and Chen Qian.

- 目标检测 #1

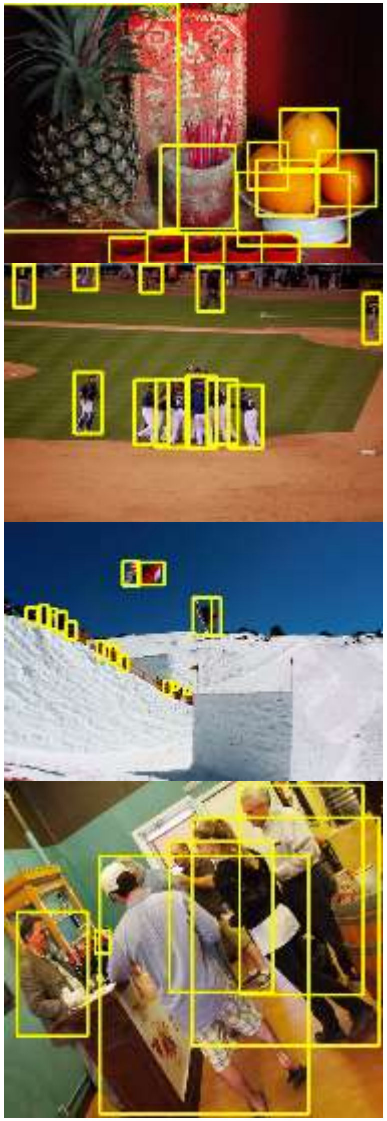
CentripetalNet: Pursuing High-quality Keypoint Pairs for Object Detection

- Results



Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage:							
Faster R-CNN w/FPN [19]	ResNet-101 [13]	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN [11]	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2
HTC [2]	ResNeXt-101	47.1	63.9	44.7	22.8	43.9	54.6
PANet(multi-scale) [22]	ResNeXt-101	47.4	67.2	51.8	30.1	**51.7**	60.0
TridentNet(multi-scale) [18]	ResNet-101-DCN	**48.4**	**69.7**	**53.5**	**31.8**	51.3	**60.3**
Single-stage anchor-based:							
SSD513 [23]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
YOLOv3 [28]	DarkNet-53	33.0	57.9	34.4	18.3	35.4	41.9
RetinaNet800 [20]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
Single-stage anchor-free:							
ExtremeNet(single-scale) [38]	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1
CornerNet511(multi-scale) [17]	Hourglass-104	42.1	57.8	45.3	20.8	44.8	56.7
FCOS [31]	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
ExtremeNet(multi-scale) [38]	Hourglass-104	43.7	60.5	47.0	24.1	46.9	57.6
CenterNet511(single-scale) [7]	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
RPDet(single-scale) [35]	ResNet-101-DCN	45.0	66.1	49.0	26.6	48.6	57.5
RPDet(multi-scale) [35]	ResNet-101-DCN	46.5	**67.4**	50.9	**30.3**	49.7	57.1
CenterNet511(multi-scale) [7]	Hourglass-104	47.0	64.5	50.7	28.9	49.9	58.9
CentripetalNet w.o/mask(single-scale)	Hourglass-104	45.8	63.0	49.3	25.0	48.2	58.7
CentripetalNet w.o/mask(multi-scale)	Hourglass-104	47.8	65.0	51.5	28.9	50.2	59.4
CentripetalNet(single-scale)	Hourglass-104	46.1	63.1	49.7	25.3	48.7	59.2
CentripetalNet(multi-scale)	Hourglass-104	**48.0**	65.1	**51.8**	29.0	**50.4**	**59.9**

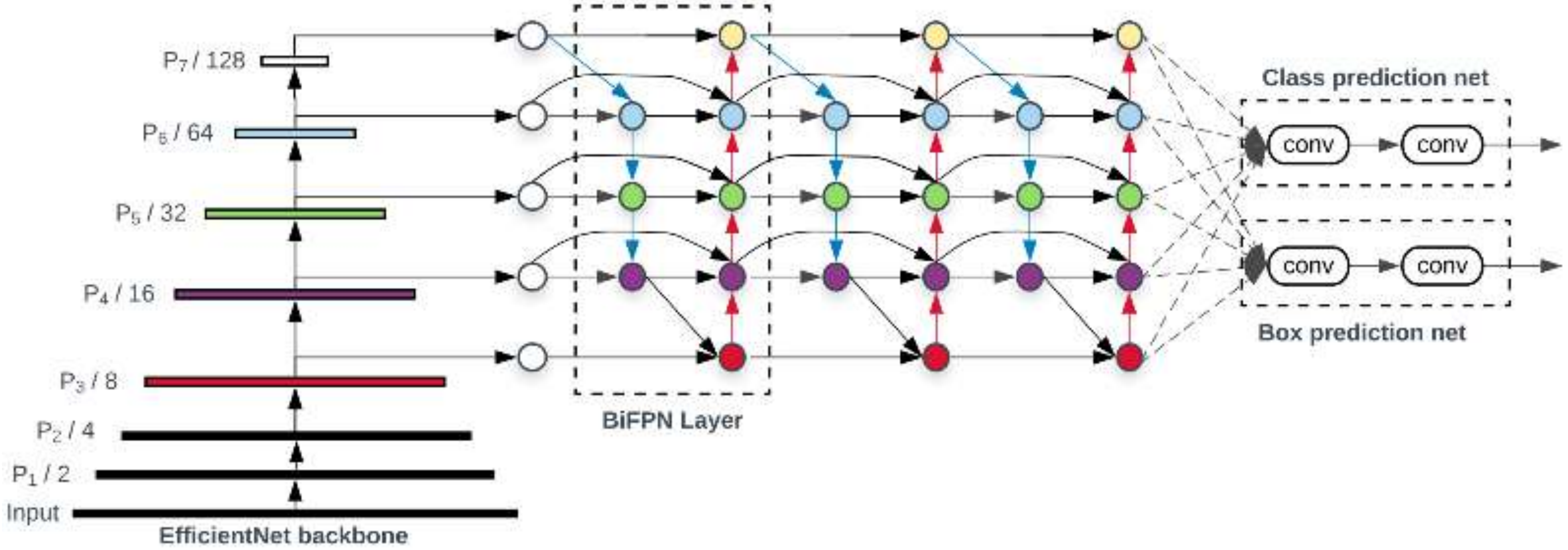
Table 1. Object detection performance comparison on MS-COCO test-dev.



- 目标检测 #2

EfficientDet: Scalable and Efficient Object Detection

- Abstract & Architecture



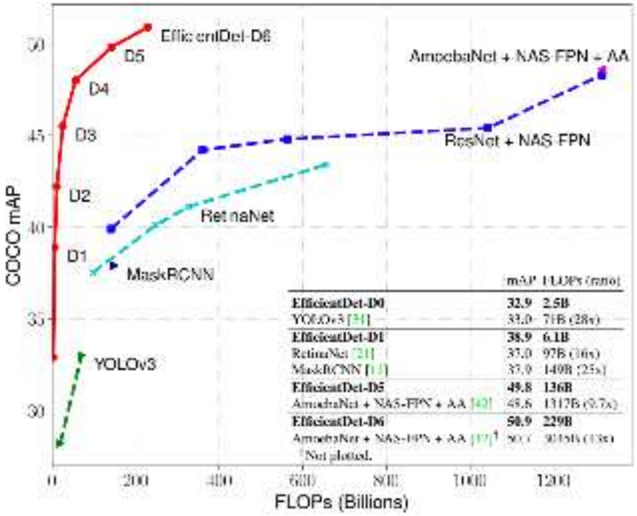
Model efficiency has become increasingly important in computer vision. In this paper, we systematically study neural network architecture design choices for object detection and propose several key optimizations to improve efficiency. First, we propose a weighted **bi-directional feature pyramid network (BiFPN)**, which allows easy and fast multi-scale feature fusion; Second, we propose a **compound scaling method** that uniformly scales the resolution, depth, and width for all backbone, feature network, and box/class prediction networks at the same time. Based on these optimizations, we have developed a new family of object detectors, called **EfficientDet**, which consistently achieve much better efficiency than prior art across a wide spectrum of resource constraints. In particular, with single model and single scale, our EfficientDet-D6 achieves state-of-the-art 50.9 mAP on COCO dataset with 52M parameters and 229B FLOPs1, being 4x smaller and using 13x fewer FLOPs yet still more accurate (+0.2% mAP) than the best previous detector.

CVPR 2020. Mingxing Tan, Ruoming Pang, and Quoc V. Le.

- 目标检测 #2

EfficientDet: Scalable and Efficient Object Detection

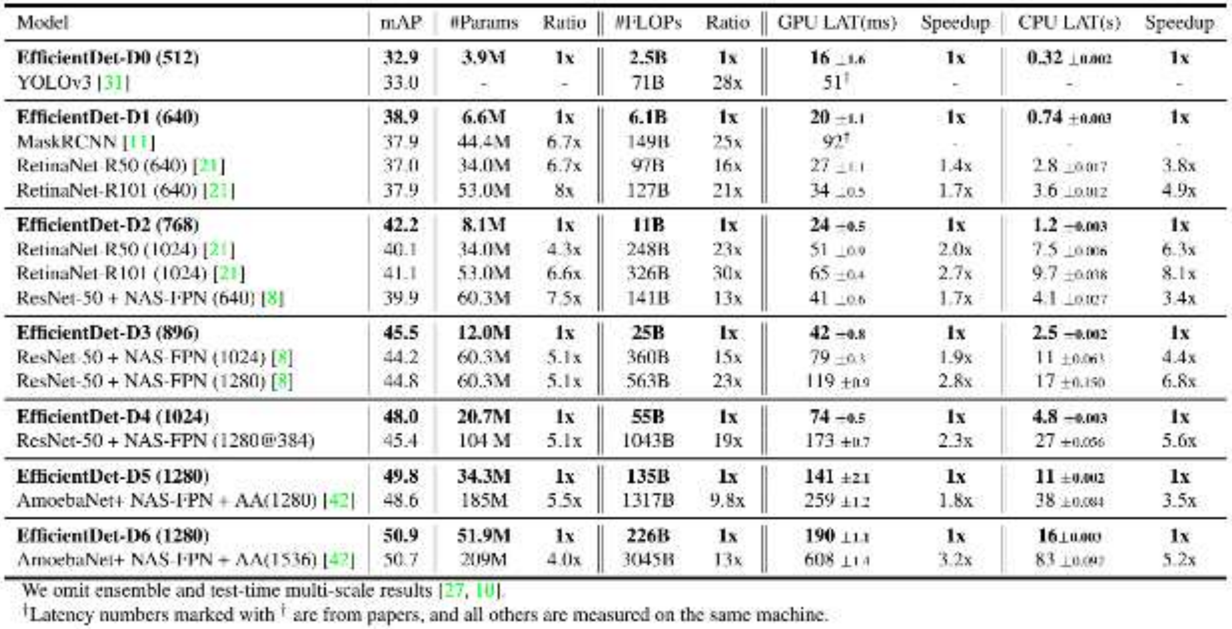
- Results & Ablation Study



	mAP	FLOPs (ratio)
EfficientDet-D0	**32.9**	**2.5B**
YOLOv3 [31]	33.0	71B (28x)
EfficientDet-D1	**38.9**	**6.1B**
RetinaNet [21]	37.0	97B (16x)
MaskRCNN [11]	37.9	149B (25x)
EfficientDet-D5	**49.8**	**135B**
AmoebaNet + NAS-FPN + AA [42]	48.6	1317B (9.7x)
EfficientDet-D6	**50.9**	**229B**
AmoebaNet + NAS-FPN + AA [42]†	50.7	3045B (13x)

†Not plotted.

Figure 1: **Model FLOPs vs COCO accuracy** – All numbers are for single-model single-scale. Our EfficientDet achieves much better accuracy with fewer computations. In particular, EfficientDet-D6 achieves new state-of-the-art 50.9% COCO mAP with 4x fewer parameters and 13x fewer FLOPs than prior art [42].

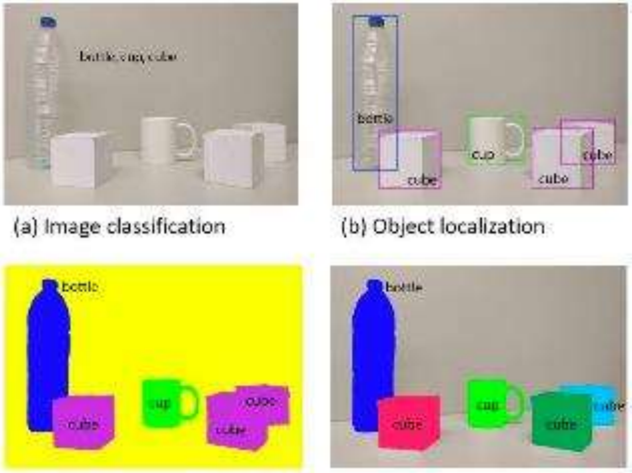


Model	mAP	#Params	Ratio	#FLOPs	Ratio	GPU LAT(ms)	Speedup	CPU LAT(s)	Speedup
EfficientDet-D0 (512)	**32.9**	**3.9M**	**1x**	**2.5B**	**1x**	**16** ±1.6	**1x**	**0.32** ±0.002	**1x**
YOLOv3 [31]	33.0	-	-	71B	28x	51†	-	-	-
EfficientDet-D1 (640)	**38.9**	**6.6M**	**1x**	**6.1B**	**1x**	**20** ±1.1	**1x**	**0.74** ±0.003	**1x**
MaskRCNN [11]	37.9	44.4M	6.7x	149B	25x	92†	-	-	-
RetinaNet-R50 (640) [21]	37.0	34.0M	6.7x	97B	16x	27 ±1.1	1.4x	2.8 ±0.017	3.8x
RetinaNet-R101 (640) [21]	37.9	53.0M	8x	127B	21x	34 ±0.5	1.7x	3.6 ±0.012	4.9x
EfficientDet-D2 (768)	**42.2**	**8.1M**	**1x**	**11B**	**1x**	**24** ±0.5	**1x**	**1.2** ±0.003	**1x**
RetinaNet-R50 (1024) [21]	40.1	34.0M	4.3x	248B	23x	51 ±0.9	2.0x	7.5 ±0.006	6.3x
RetinaNet-R101 (1024) [21]	41.1	53.0M	6.6x	326B	30x	65 ±0.4	2.7x	9.7 ±0.008	8.1x
ResNet-50 + NAS-FPN (640) [8]	39.9	60.3M	7.5x	141B	13x	41 ±0.6	1.7x	4.1 ±0.027	3.4x
EfficientDet-D3 (896)	**45.5**	**12.0M**	**1x**	**25B**	**1x**	**42** ±0.8	**1x**	**2.5** ±0.002	**1x**
ResNet-50 + NAS-FPN (1024) [8]	44.2	60.3M	5.1x	360B	15x	79 ±0.3	1.9x	11 ±0.063	4.4x
ResNet-50 + NAS-FPN (1280) [8]	44.8	60.3M	5.1x	563B	23x	119 ±0.9	2.8x	17 ±0.150	6.8x
EfficientDet-D4 (1024)	**48.0**	**20.7M**	**1x**	**55B**	**1x**	**74** ±0.5	**1x**	**4.8** ±0.003	**1x**
ResNet-50 + NAS-FPN (1280@384)	45.4	104 M	5.1x	1043B	19x	173 ±0.7	2.3x	27 ±0.056	5.6x
EfficientDet-D5 (1280)	**49.8**	**34.3M**	**1x**	**135B**	**1x**	**141** ±2.1	**1x**	**11** ±0.002	**1x**
AmoebaNet+ NAS-FPN + AA(1280) [42]	48.6	185M	5.5x	1317B	9.8x	259 ±1.2	1.8x	38 ±0.084	3.5x
EfficientDet-D6 (1280)	**50.9**	**51.9M**	**1x**	**226B**	**1x**	**190** ±1.1	**1x**	**16** ±0.003	**1x**
AmoebaNet+ NAS-FPN + AA(1536) [42]	50.7	209M	4.0x	3045B	13x	608 ±1.4	3.2x	83 ±0.092	5.2x

We omit ensemble and test-time multi-scale results [27, 10].
†Latency numbers marked with † are from papers, and all others are measured on the same machine.

Table 2: **EfficientDet performance on COCO** [22] – Results are for single-model single-scale. #Params and #FLOPs denote the number of parameters and multiply-adds. LAT denotes inference latency with batch size 1. AA denotes auto-augmentation [42]. We group models together if they have similar accuracy, and compare the ratio or speedup between EfficientDet and other detectors in each group.

语义分割



(a) Image classification (b) Object localization

(c) Semantic segmentation (d) Instance segmentation

语义分割：把图像中的每个像素都划分到某一个类别上	
常用数据集	Cityscapes、COCO、Pascal VOC、KITTI、PASCAL-Context等
常用指标	AP，预测正确的像素占总像素的比例
	mAP，每个类被正确分类像素数的比例
	mIoU，平均交并比
代表算法	1.Trevor Darrell(UC Berkeley)——FCN
	2.Facebook AI Research——Mask R-CNN、RetinaNet
	3.Google AI——DeepLab系列
	4.Thomas Brox(德国弗莱堡大学)——U-Net
	5.Fisher Yu(Princeton University)——Dilated Convolution、DRN
	6.Hanqing Lu(中科院自动化研究所)——DANet

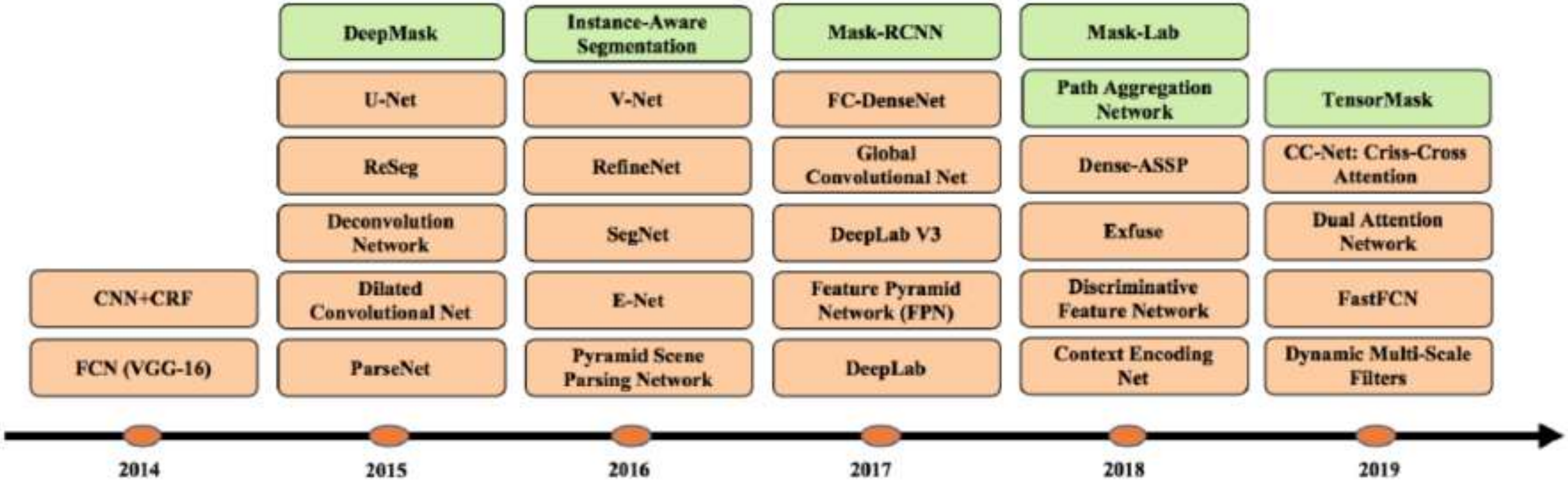
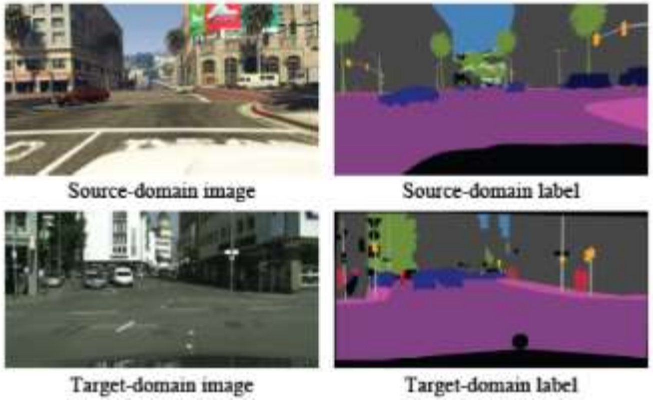


Fig. 41. The timeline of DL-based segmentation algorithms for 2D images. Orange and green blocks refer to semantic, and instance segmentation algorithms respectively.

- **Domain Adaptation Semantic Segmentation**



Source-domain image

Source-domain label

Target-domain image

Target-domain label

- Synthia → CityScapes
- GTA-5 → CityScapes

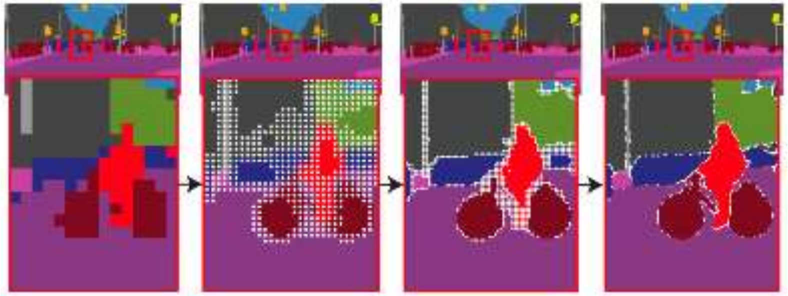
arXiv 2020. Image Segmentation Using Deep Learning: A Survey

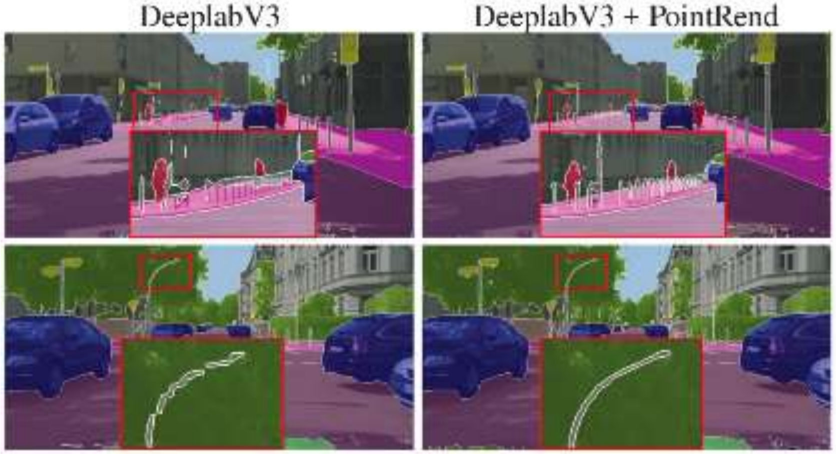
- 语义分割 #1

PointRend: Image Segmentation as Rendering

- Abstract, Adaptive Subdivison Step & Results

Example of one **adaptive subdivision step**. A prediction is upsampled by 2× using bilinear interpolation. Then, PointRend makes prediction for the N most ambiguous points (white dots) to recover detail on the finer grid. This process is repeated until the desired resolution is achieved.





method	output resolution	mIoU
DeeplabV3-OS-16	64×128	77.2
DeeplabV3-OS-8	128×256	77.8 (+0.6)
DeeplabV3-OS-16 + PointRend	1024×2048	**78.4** (+1.2)

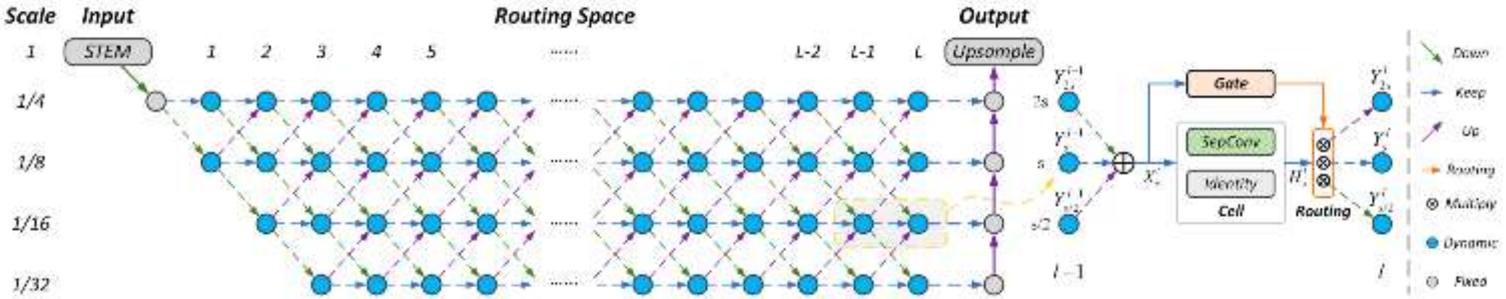
We present a new method for efficient high-quality image segmentation of objects and scenes. By analogizing classical computer graphics methods for efficient rendering with over- and undersampling challenges faced in pixel labeling tasks, we develop a unique perspective of image segmentation as a rendering problem. From this vantage, we present the **PointRend (Point-based Rendering)** neural network module: a module that performs point-based segmentation predictions at adaptively selected locations based on an iterative subdivision algorithm. PointRend can be flexibly applied to both instance and semantic segmentation tasks by building on top of existing state-of-the-art models. While many concrete implementations of the general idea are possible, we show that a simple design already achieves excellent results. PointRend's efficiency enables output resolutions that are otherwise impractical in terms of memory or computation compared to existing approaches.

CVPR 2020. Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick.

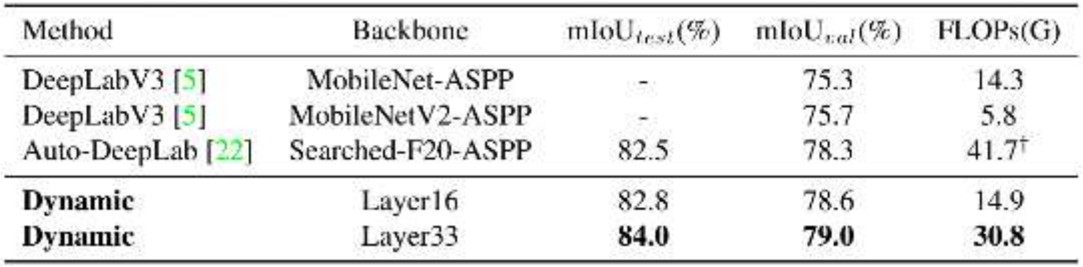
- 语义分割 #2

Learning Dynamic Routing for Semantic Segmentation

- Abstract, Framework & Results



Recently, numerous handcrafted and searched networks have been applied for semantic segmentation. However, previous works intend to handle inputs with various scales in pre-defined static architectures, such as FCN, U-Net, and DeepLab series. This paper studies a conceptually new method to alleviate the scale variance in semantic representation, named **dynamic routing**. The proposed framework generates data-dependent routes, adapting to the scale distribution of each image. To this end, a differentiable gating function, called **soft conditional gate**, is proposed to select scale transform paths on the fly. In addition, the computational cost can be further reduced in an end-to-end manner by giving budget constraints to the gating function. We further relax the network level routing space to support multipath propagations and skip-connections in each forward, bringing substantial network capacity. To demonstrate the superiority of the dynamic property, we compare with several static architectures, which can be modeled as special cases in the routing space.



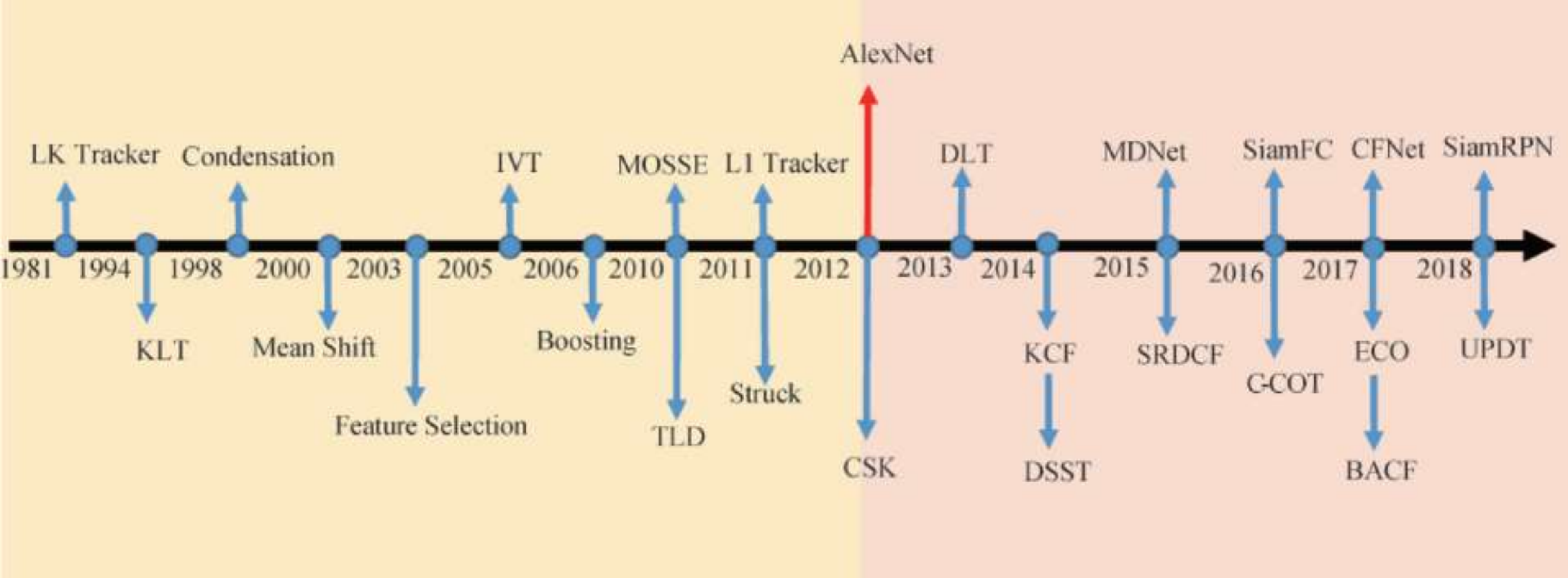
Method	Backbone	mIoU_{test}(%)	mIoU_{val}(%)	FLOPs(G)
DeepLabV3 [5]	MobileNet-ASPP	-	75.3	14.3
DeepLabV3 [5]	MobileNetV2-ASPP	-	75.7	5.8
Auto-DeepLab [22]	Searched-F20-ASPP	82.5	78.3	$41.7^{\dagger}$
Dynamic	Layer16	82.8	78.6	14.9
Dynamic	Layer33	**84.0**	**79.0**	**30.8**

CVPR 2020. Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun.

视频跟踪

视频跟踪：在连续的视频帧中定位某一物体

常用数据集	VOT、MOT、GOT、OTB、LaSOT等
常用指标	Precision plot，预测与标注目标中心点距离小于给定阈值的帧的百分比
	Success Plot，总的成功的帧占所有帧的百分比
	EAO, Expect Average Overlaprate, 平均重叠率
	A, Accuracy, 准确率；R, Robustness, 鲁棒性
代表算法	1.Joao F. Henriques, Luca Bertinetto(University of Oxford)——SiamFC、Staple、Learnet、CSK、KCF/DCF、CFNet
	2.Martin Danelljan(林雪平大学)——CN、DSST、SRDCF、DeepSRDCF、SRDCFdecon、CCOT、ECO
	3.Naiyan Wang (Winsty)——DLT、SO-DLT、UDVTS
	4.Tencent AI Lab——UDT
	5.Hua Yang(上交大)——DMAN



- **基于生成式模型的方法：**生成式模型提取目标特征构建表观模型，在图像中搜索与模型最匹配的区域作为跟踪结果。代表方法：LK光流法、L1跟踪器等。
- **基于判别式模型的方法：**与生成式模型不同的是，判别式模型同时考虑了目标和背景信息。判别式模型将跟踪问题看做分类（TLD，tracking learning detection）或者回归（核相关滤波）问题，目的是寻找一个判别函数，将目标从背景中分离出来，从而实现对目标的跟踪。

- **基于深度学习的方法**

  1. 基于预训练深度特征的跟踪：HCF、HDT、C-COT、UPDT
  2. 基于离线训练特征的跟踪：MDNet、SiamFC、SINT

中国图象图形学报 2019. 深度学习的目标跟踪算法综述

- **用于视频跟踪的主干网络 (Network Backbone)**

  1. 深度判别式模型：CNN（AlexNet、VGGNet、GoogLeNet、ResNet、DenseNet），RNN（LSTM、GRU、ConvLSTM）…
  2. 深度生成式模型：GAN（DCGAN、WGAN、WGAN-GP）、AE（AE、VAE）…
  3. 其他深度学习模型：强化学习(RL)、元学习（Meta Learning）…

- **按照网络结构分类的视频跟踪方法**

  1. 基于卷积神经网络的深度目标跟踪方法（DNT、CNT）
  2. 基于递归神经网络的深度目标跟踪方法（SANet、MemTrack）
  3. 基于生成式对抗网络的深度目标跟踪方法（VITAL、SINT++、ADT）
  4. 基于自编码器的深度目标跟踪方法（TRACA、EDCF）

- **按照网络功能分类的视频跟踪方法**

  1. 基于相关滤波的深度目标跟踪方法（HCF、CFNet、FlowTrack、C-COT、ECO、DRT）
  2. 基于分类网络的深度目标跟踪方法（MDNet、VITAL、ADNet）
  3. 基于回归网络的深度目标跟踪方法（DSLT、GOTURN）

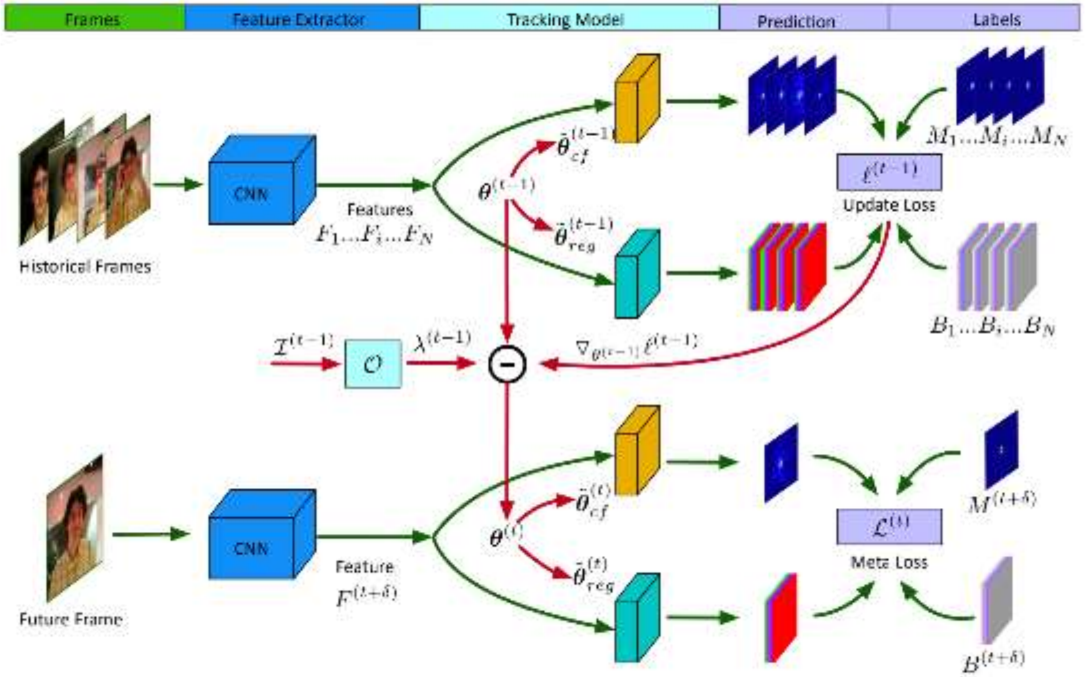
- **按照网络训练分类的视频跟踪方法**

  1. 基于预训练网络的深度目标跟踪方法（HDT、DeepSRDCF）
  2. 基于在线微调网络的深度目标跟踪方法（MDNet、CREST、ADNet）
  3. 基于离线训练网络的深度目标跟踪方法（SiamFC、GOTURN）

- 视频跟踪 #1

ROAM: Recurrently Optimizing Tracking Model

- Abstract & Pipeline



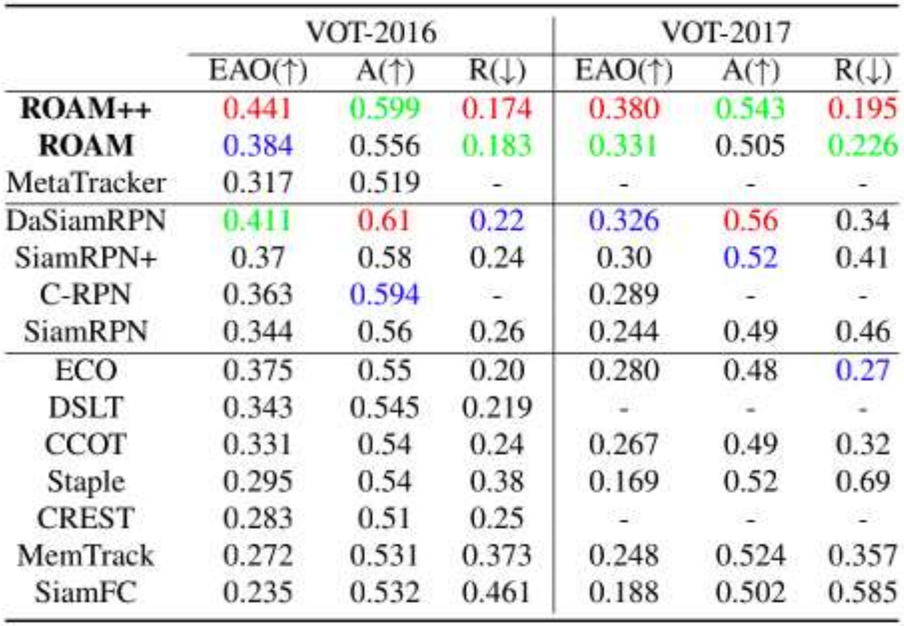
In this paper, we design a tracking model consisting of response generation and bounding box regression, where the first component produces a heat map to indicate the presence of the object at different positions and the second part regresses the relative bounding box shifts to anchors mounted on sliding-window locations. Thanks to the resizable convolutional filters used in both components to adapt to the shape changes of objects, our tracking model does not need to enumerate different sized anchors, thus saving model parameters. To effectively adapt the model to appearance variations, we propose to offline train a recurrent neural optimizer to update tracking model in a meta-learning setting, which can converge the model in a few gradient steps. This improves the convergence speed of updating the tracking model while achieving better performance.

CVPR 2020. Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B. Chan.

- 视频跟踪 #1

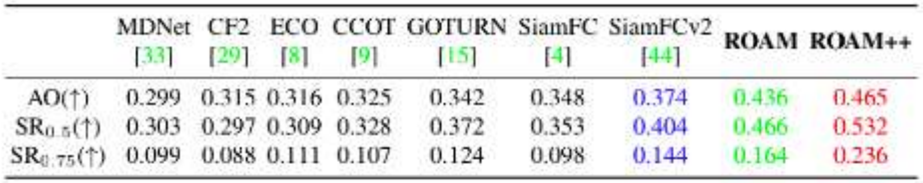
ROAM: Recurrently Optimizing Tracking Model

- Results



	VOT-2016			VOT-2017		
	EAO(↑)	A(↑)	R(↓)	EAO(↑)	A(↑)	R(↓)
ROAM++	0.441	0.599	0.174	0.380	0.543	0.195
ROAM	0.384	0.556	0.183	0.331	0.505	0.226
MetaTracker	0.317	0.519	-	-	-	-
DaSiamRPN	0.411	0.61	0.22	0.326	0.56	0.34
SiamRPN+	0.37	0.58	0.24	0.30	0.52	0.41
C-RPN	0.363	0.594	-	0.289	-	-
SiamRPN	0.344	0.56	0.26	0.244	0.49	0.46
ECO	0.375	0.55	0.20	0.280	0.48	0.27
DSLT	0.343	0.545	0.219	-	-	-
CCOT	0.331	0.54	0.24	0.267	0.49	0.32
Staple	0.295	0.54	0.38	0.169	0.52	0.69
CREST	0.283	0.51	0.25	-	-	-
MemTrack	0.272	0.531	0.373	0.248	0.524	0.357
SiamFC	0.235	0.532	0.461	0.188	0.502	0.585

Table 1: Results on VOT-2016/2017. The evaluation metrics are expected average overlap (EAO), accuracy value (A), robustness value (R). The top 3 performing trackers are colored with red, green and blue respectively.



	MDNet [33]	CF2 [29]	ECO [8]	CCOT [9]	GOTURN [15]	SiamFC [4]	SiamFCv2 [44]	**ROAM**	**ROAM++**
AO(↑)	0.299	0.315	0.316	0.325	0.342	0.348	0.374	0.436	0.465
$SR_{0.5}$(↑)	0.303	0.297	0.309	0.328	0.372	0.353	0.404	0.466	0.532
$SR_{0.75}$(↑)	0.099	0.088	0.111	0.107	0.124	0.098	0.144	0.164	0.236

Table 2: Results on GOT-10k. The evaluation metrics include average overlap (AO), success rate at 0.5 overlap threshod. ($SR_{0.5}$), success rate at 0.75 overlap threshod. ($SR_{0.75}$). The top three performing trackers are colored with red, green and blue respectively.

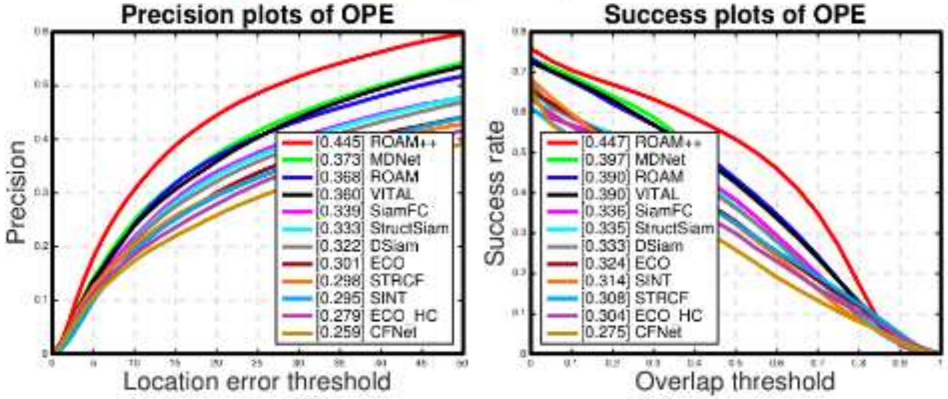


Figure 4: Precision and success plot on LaSOT test dataset

- 视频跟踪 #2

A Unified Object Motion and Affinity Model for Online Multi-Object Tracking

- Abstract & UMA Model

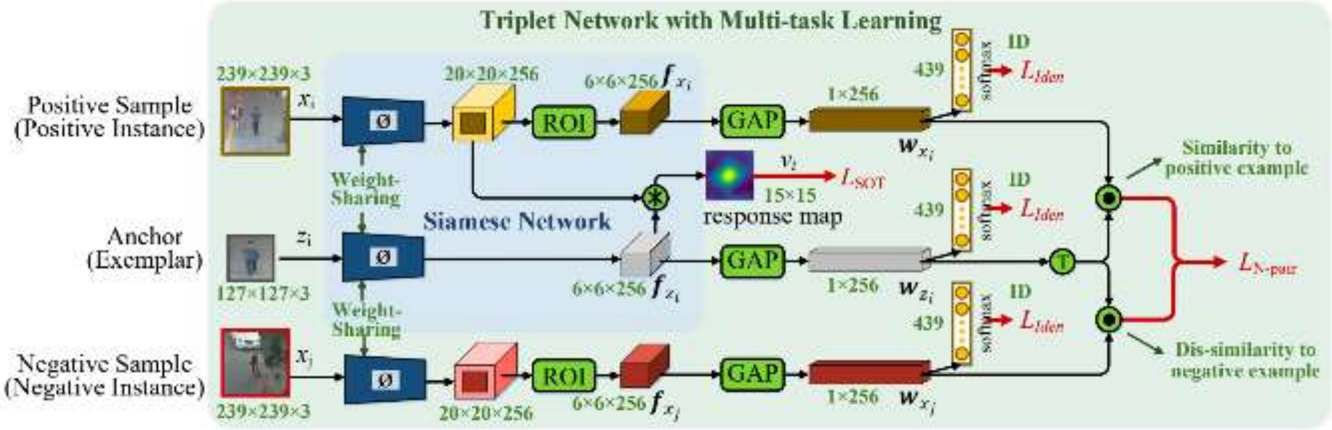


Figure 2: **Illustration of our proposed UMA model**, which is built upon a triplet architecture with multi-task learning. UMA simultaneously learns two tasks: SOT based object motion prediction and affinity-dependent ranking, producing a strong feature which is applicable to both the tracklet generation as well as the affinity measure phases.

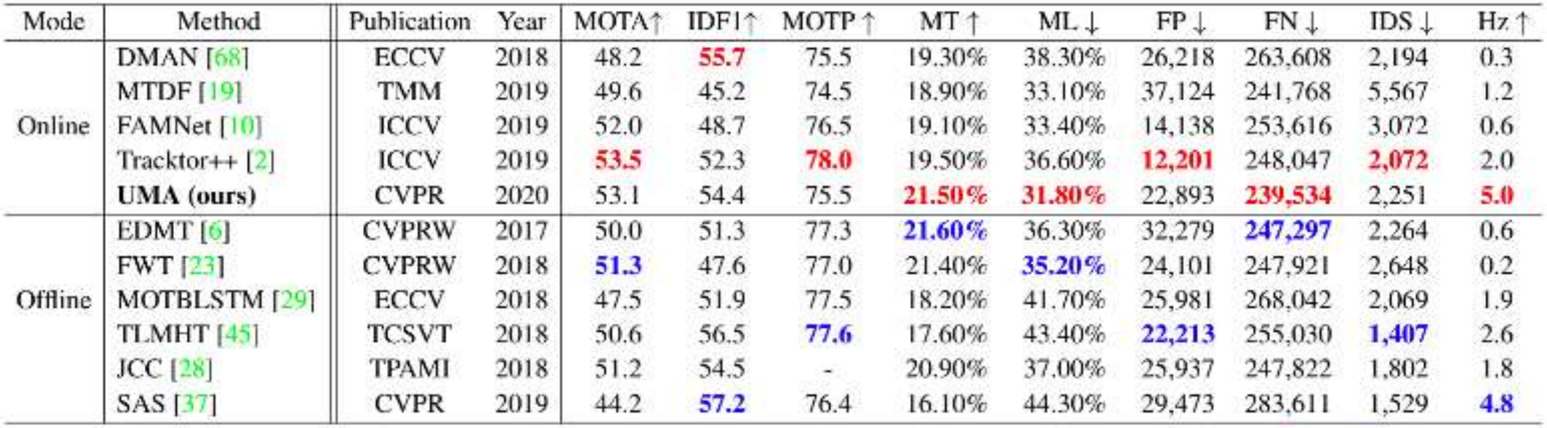
Current popular online multi-object tracking (MOT) solutions apply single object trackers (SOTs) to capture object motions, while often requiring an extra affinity network to associate objects, especially for the occluded ones. This brings extra computational overhead due to repetitive feature extraction for SOT and affinity computation. Mean_x0002_while, the model size of the sophisticated affinity network is usually non-trivial. In this paper, we propose a novel MOT framework that unifies object motion and affinity model into a single network, named **UMA**, in order to learn a compact feature that is discriminative for both object motion and affinity measure. In particular, UMA integrates single object tracking and metric learning into a unified triplet network by means of multi-task learning. Such design brings advantages of improved computation efficiency, low memory requirement and simplified training procedure. In addition, we equip our model with a task-specific attention module, which is used to boost task-aware feature learning. The proposed UMA can be easily trained end-to-end, and is elegant – requiring only one training stage.

CVPR 2020. Junbo Yin, Wenguan Wang*, Qinghao Meng, Ruigang Yang, and Jianbing Shen.

- 视频跟踪 #2

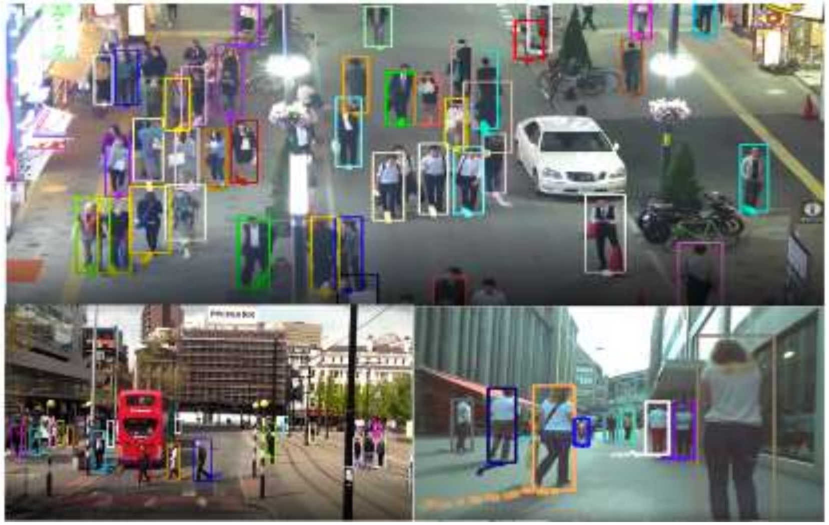
A Unified Object Motion and Affinity Model for Online Multi-Object Tracking

- Results



Mode	Method	Publication	Year	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Hz↑
Online	DMAN [68]	ECCV	2018	48.2	**55.7**	75.5	19.30%	38.30%	26,218	263,608	2,194	0.3
	MTDF [19]	TMM	2019	49.6	45.2	74.5	18.90%	33.10%	37,124	241,768	5,567	1.2
	FAMNet [10]	ICCV	2019	52.0	48.7	76.5	19.10%	33.40%	14,138	253,616	3,072	0.6
	Tracktor++ [2]	ICCV	2019	**53.5**	52.3	**78.0**	19.50%	36.60%	**12,201**	248,047	**2,072**	2.0
	UMA (ours)	CVPR	2020	53.1	54.4	75.5	**21.50%**	**31.80%**	22,893	**239,534**	2,251	**5.0**
Offline	EDMT [6]	CVPRW	2017	50.0	51.3	77.3	**21.60%**	36.30%	32,279	**247,297**	2,264	0.6
	FWT [23]	CVPRW	2018	**51.3**	47.6	77.0	21.40%	**35.20%**	24,101	247,921	2,648	0.2
	MOTBLSTM [29]	ECCV	2018	47.5	51.9	77.5	18.20%	41.70%	25,981	268,042	2,069	1.9
	TLMHT [45]	TCSVT	2018	50.6	56.5	**77.6**	17.60%	43.40%	**22,213**	255,030	**1,407**	2.6
	JCC [28]	TPAMI	2018	51.2	54.5	-	20.90%	37.00%	25,937	247,822	1,802	1.8
	SAS [37]	CVPR	2019	44.2	**57.2**	76.4	16.10%	44.30%	29,473	283,611	1,529	**4.8**

Table 2: **Quantitative results on MOT17.** The best scores of online and offline MOT methods are marked in **red** and **blue**, respectively.



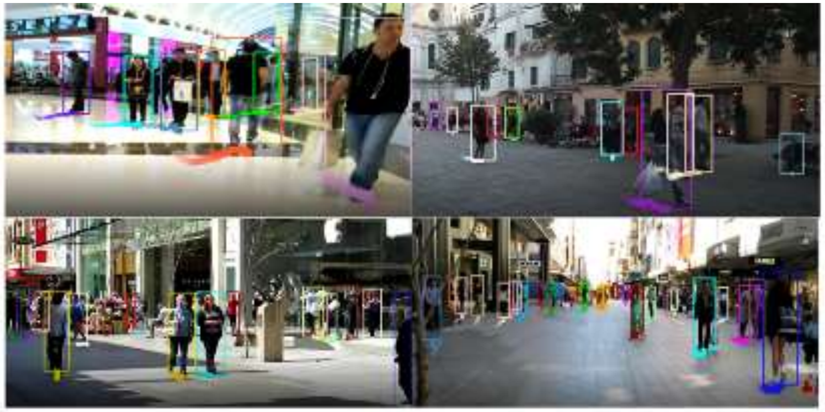
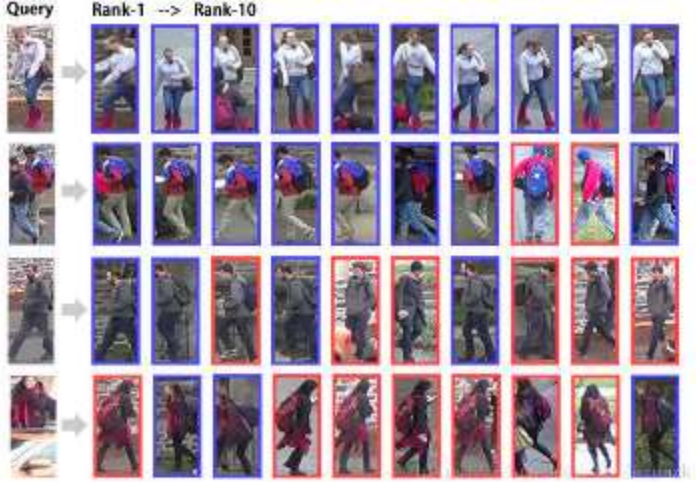


Figure 7: **Qualitative tracking results on the test sequences of MOT17 benchmark.** The color of each bounding box indicates the target identity. The dotted line under each bounding box denotes the recent tracklet of each target.

ReID



ReID：给定一个监控行人图像，检索跨设备下的该行人图像	
常用数据集	CUHK、Market1501、DukeMTMC-reID、MSMT17等
常用指标	mAP，检索的人在数据库中所有正确的图片排在排序列表前面的程度
	Rank-n，搜索结果中最靠前（置信度最高）的n张图有正确结果的概率
	CMC, Cumulative Match Characteristic curve，累积匹配曲线
	F-score , recall和precision的调和平均数 2 * P * R / (P + R)
代表算法	1.Wen Gao, Qi Tian——PTGAN、MTL-LORAE, ScalablePRID
	2.Stan Z. Li(CASIA)——LOMO、DML
	3.Weishi Zheng(中山大学)——MAR、CAMEL
	4.Shaogang Gong(QMUL)——TJ-AIDL、HA-CNN
	5.旷视Face++——AlignedReID

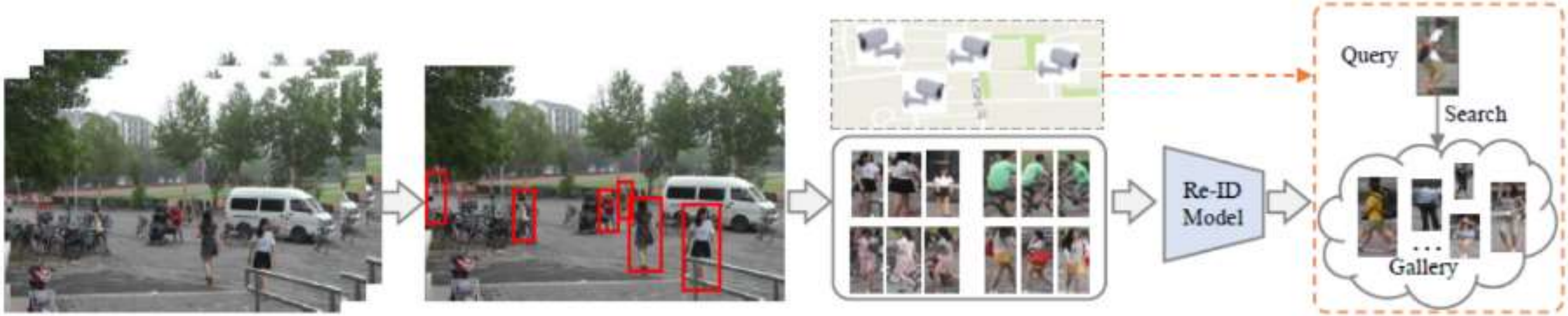


Fig. 1: The flow of designing a practical person Re-ID system, including five main steps: 1) *Raw Data Collection*, (2) *Bounding Box Generation*, 3) *Training Data Annotation*, 4) *Model Training* and 5) *Pedestrian Retrieval*.

- **Re-ID的五大步骤**

  1. 数据收集
  2. 包围框生成
  3. 训练数据标注
  4. 模型训练
  5. 行人检索

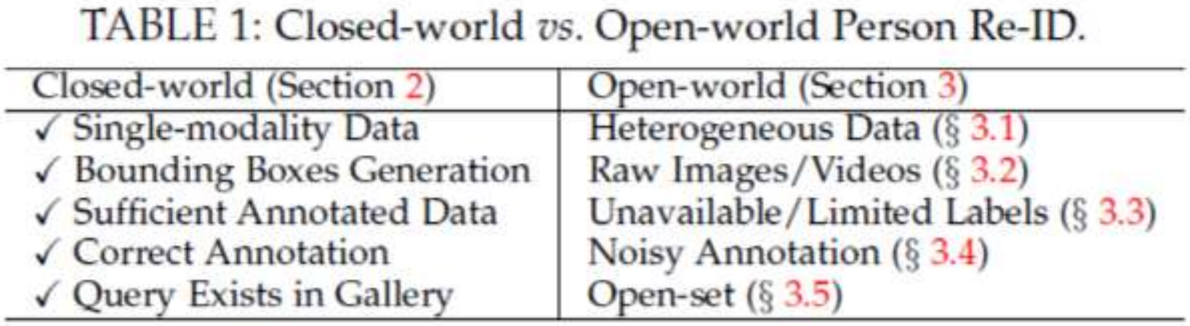


TABLE 1: Closed-world *vs.* Open-world Person Re-ID.

Closed-world (Section 2)	Open-world (Section 3)
✓ Single-modality Data	Heterogeneous Data (§ 3.1)
✓ Bounding Boxes Generation	Raw Images/Videos (§ 3.2)
✓ Sufficient Annotated Data	Unavailable/Limited Labels (§ 3.3)
✓ Correct Annotation	Noisy Annotation (§ 3.4)
✓ Query Exists in Gallery	Open-set (§ 3.5)

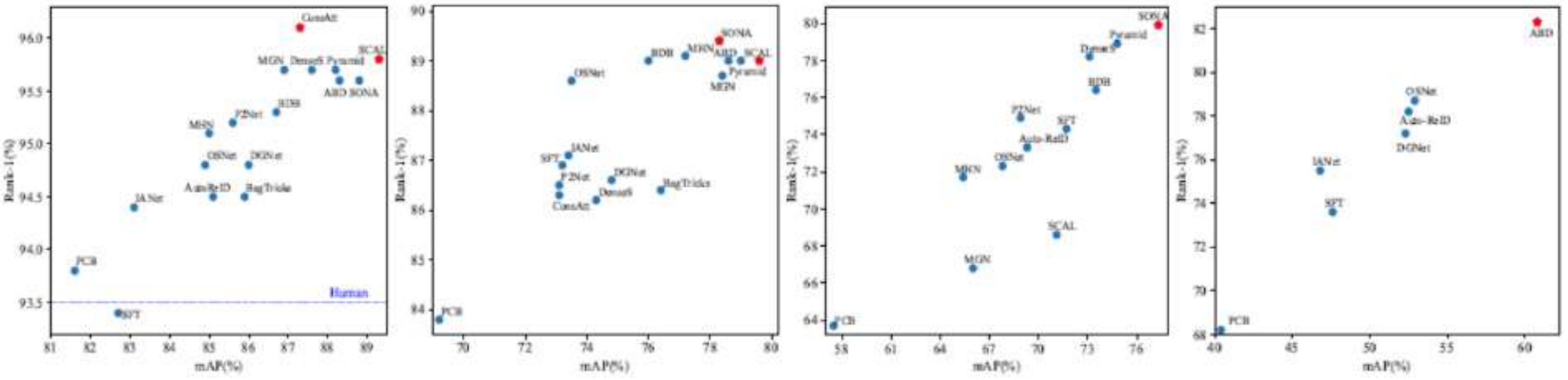
- Closed-world Person Re-identification

  ✓ Person appearances are captured by single-modality visible cameras, either by image or video
  ✓ The persons are represented by bounding boxes, where most of the bounding box area belongs the same identity
  ✓ The training has enough annotated training data for supervised discriminative Re-ID model learning
  ✓ The annotations are generally correct
  ✓ The query person must appear in the gallery set

arXiv 2020. Deep Learning for Person Re-identification

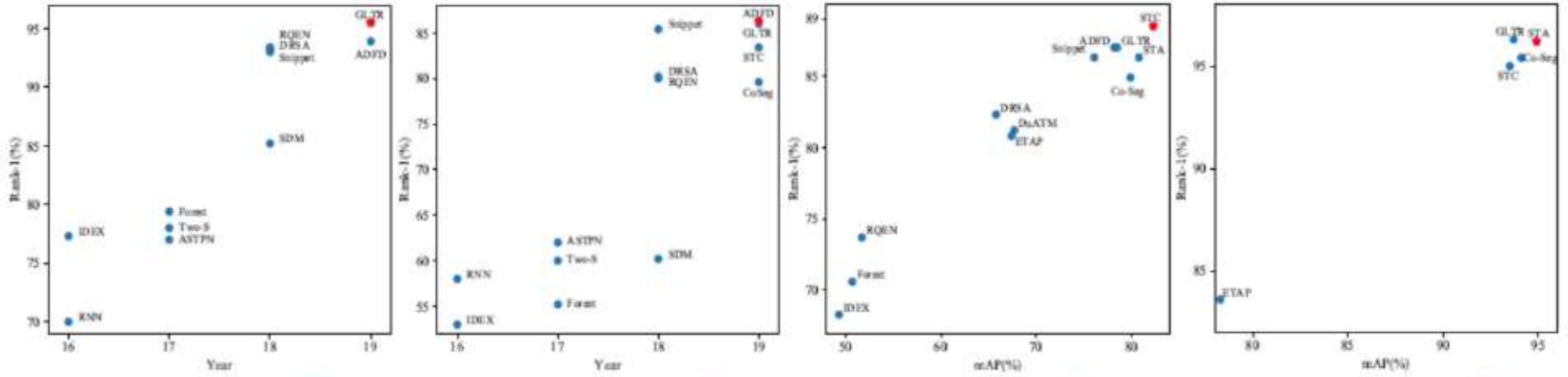
- Three main components of a standard closed-world Re-ID system

  1. Feature Representation Learning, 2. Deep Metric Learning, 3. Ranking Optimization



(a) SOTA on Market-1501 [5] (b) SOTA on DukeMTMC [33] (c) SOTA on CUHK03 [34] (d) SOTA on MSMT17 [35]

Fig. 5: State-of-the-arts (SOTA) on four widely used image-based person Re-ID datasets. Both the Rank-1 accuracy (%) and mAP value (%) are reported. For CUHK03 [34], the detected data under the setting [49] is reported. For Market-1501, the single query setting is used, and the human-level rank-1 accuracy (93.5%) is also reported [173]. The best result is highlighted with a red star.



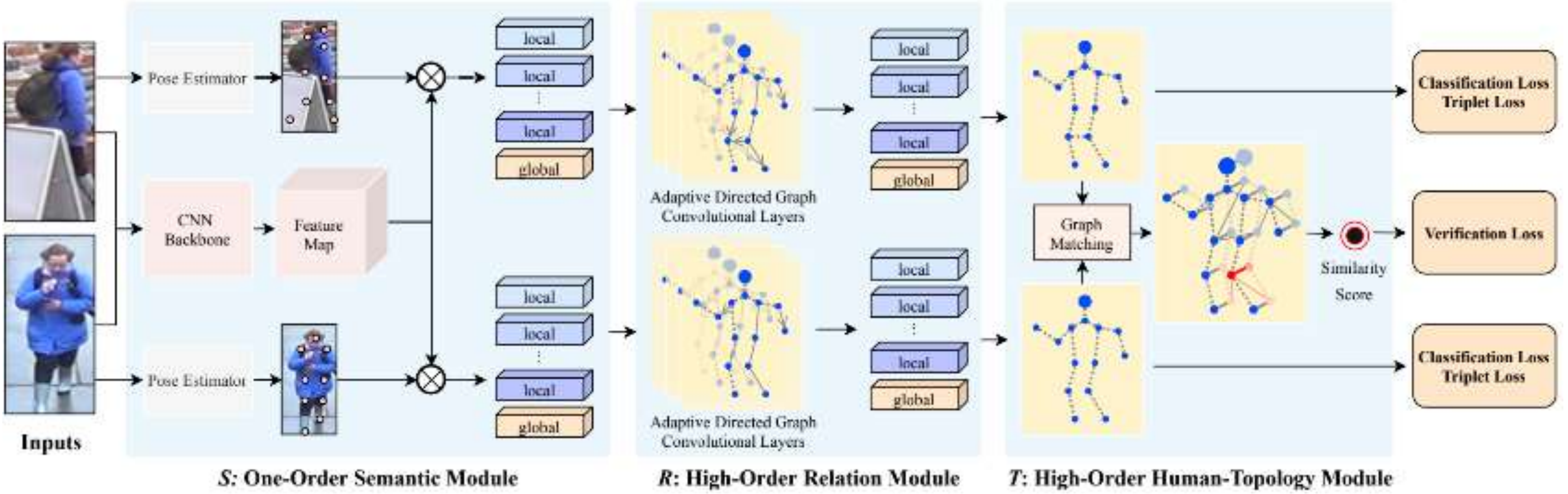
(a) SOTA on PRID-2011 [126] (b) SOTA on iLIDS-VID [6] (c) SOTA on MARS [7] (d) SOTA on Duke-Video [128]

Fig. 6: State-of-the-arts (SOTA) on four widely used video-based person Re-ID datasets. The Rank-1 accuracies (%) over years are reported. mAP values (%) on MARS [7] and Duke-Video [128] are reported. For Duke-Video, we refer to the settings in [128]. The best result is highlighted with a red star.

- ReID #1

High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification

- Abstract & Overall Framework


Inputs
Pose Estimator
CNN Backbone
Feature Map
Pose Estimator
local
local
local
global
local
local
local
global
Adaptive Directed Graph Convolutional Layers
Adaptive Directed Graph Convolutional Layers
local
local
local
global
local
local
local
global
Graph Matching
Similarity Score
Classification Loss Triplet Loss
Verification Loss
Classification Loss Triplet Loss
S: One-Order Semantic Module
R: High-Order Relation Module
T: High-Order Human-Topology Module

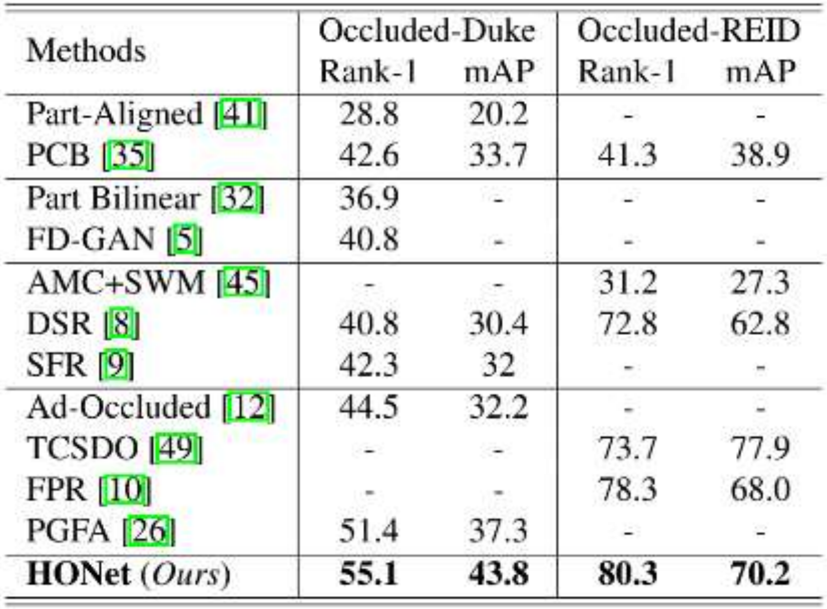
Occluded person re-identification (ReID) aims to match occluded person images to holistic ones across disjoint cameras. In this paper, we propose a novel framework by learning high-order relation and topology information for discriminative features and robust alignment. At first, we use a CNN backbone and a key-points estimation model to extract semantic local features. Even so, occluded images still suffer from occlusion and outliers. Then, we view the local features of an image as nodes of a graph and propose an **adaptive direction graph convolutional (ADGC)** layer to pass relation information between nodes. The proposed ADGC layer can automatically suppress the message passing of meaningless features by dynamically learning direction and degree of linkage. When aligning two groups of local features from two images, we view it as a graph matching problem and propose a **cross-graph embedded-alignment (CGEA)** layer to jointly learn and embed topology information to local features, and straightly predict similarity score. The proposed CGEA layer not only take full use of alignment learned by graph matching but also replace sensitive one-to-one matching with a robust soft one.

CVPR 2020.Guan’an Wang,Shuo Yang,Huanyu Liu,Zhicheng Wang,Yang Yang,Shuliang Wang,Gang Yu,Erjin Zhou,and Jian Sun.

- ReID #1

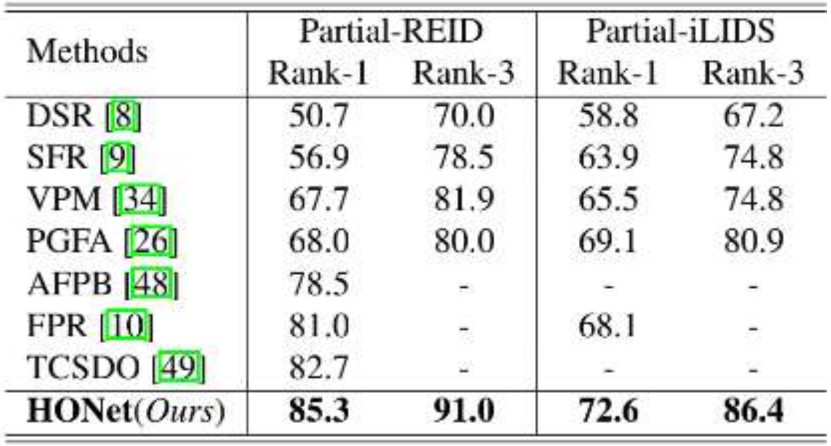
High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification

- Results



Methods	Occluded-Duke		Occluded-REID	
	Rank-1	mAP	Rank-1	mAP
Part-Aligned [41]	28.8	20.2	-	-
PCB [35]	42.6	33.7	41.3	38.9
Part Bilinear [32]	36.9	-	-	-
FD-GAN [5]	40.8	-	-	-
AMC+SWM [45]	-	-	31.2	27.3
DSR [8]	40.8	30.4	72.8	62.8
SFR [9]	42.3	32	-	-
Ad-Occluded [12]	44.5	32.2	-	-
TCSDO [49]	-	-	73.7	77.9
FPR [10]	-	-	78.3	68.0
PGFA [26]	51.4	37.3	-	-
HONet (*Ours*)	**55.1**	**43.8**	**80.3**	**70.2**

Table 2. Comparison with state-of-the-arts on two occluded datasets, *i.e.* Occluded-Duke [26] and Occluded-REID [48].



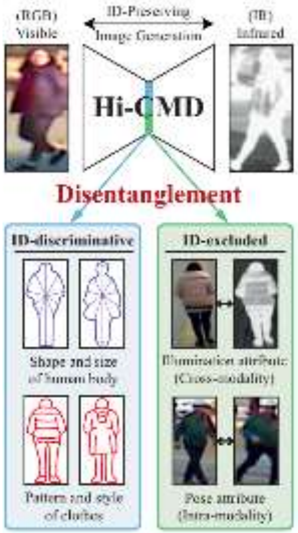
Methods	Partial-REID		Partial-iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
DSR [8]	50.7	70.0	58.8	67.2
SFR [9]	56.9	78.5	63.9	74.8
VPM [34]	67.7	81.9	65.5	74.8
PGFA [26]	68.0	80.0	69.1	80.9
AFPB [48]	78.5	-	-	-
FPR [10]	81.0	-	68.1	-
TCSDO [49]	82.7	-	-	-
HONet(*Ours*)	**85.3**	**91.0**	**72.6**	**86.4**

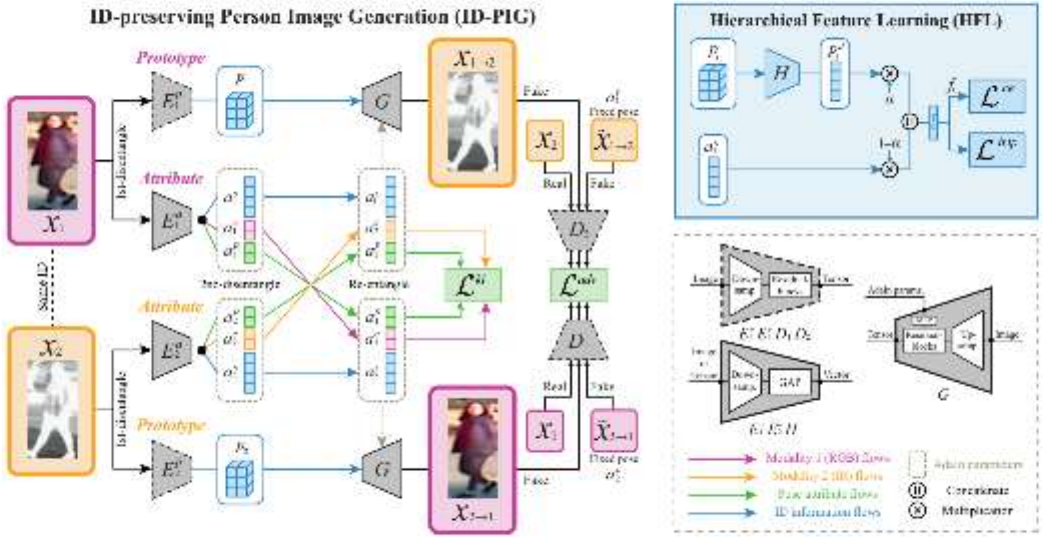
Table 3. Comparison with state-of-the-arts on two partial datasets, *i.e.* Partial-REID [45] and Partial-iLIDS [8] datasets. Our method achieves best performance on the two partial datasets.

- ReID #2

Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification

- Abstract, Concept & Framework


(RGB) Visible
ID-Preserving Image Generation
(IR) Infrared
Hi-CMD
Disentanglement
ID-discriminative
ID-excluded
Shape and size of human body
Illumination attribute (Cross-modality)
Pattern and style of clothes
Pose attribute (Intra-modality)



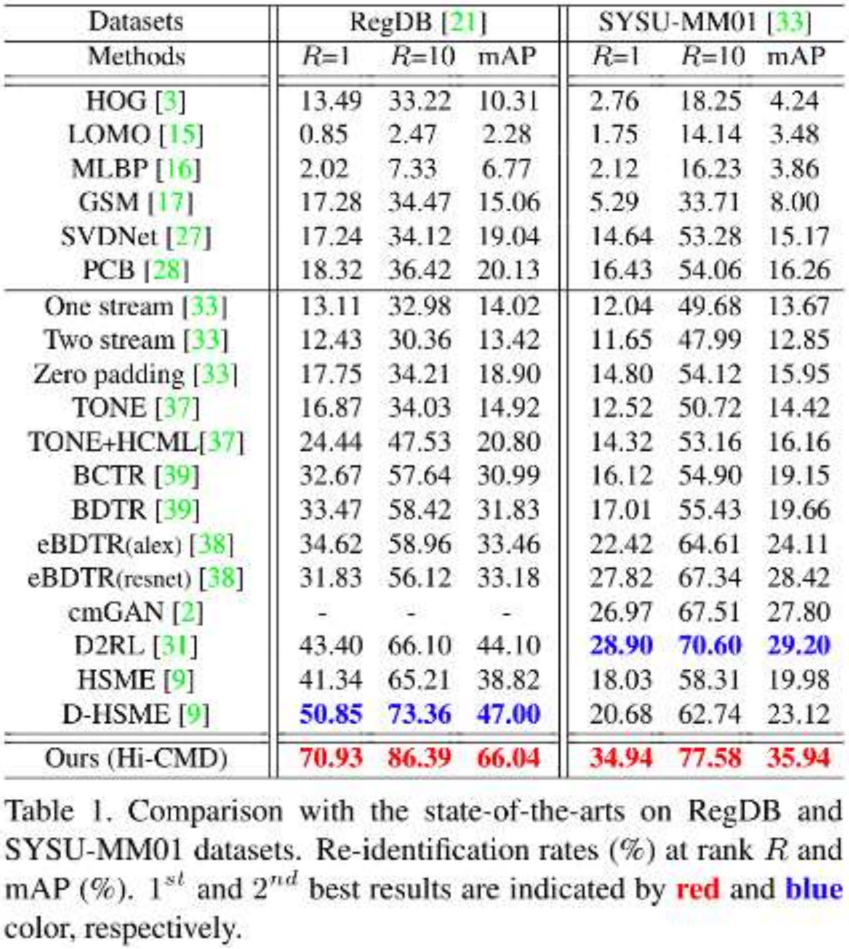
Visible-infrared person re-identification (VI-ReID) is an important task in night-time surveillance applications, since visible cameras are difficult to capture valid appearance information under poor illumination conditions. Compared to traditional person re-identification that handles only the intra-modality discrepancy, VI-ReID suffers from additional cross-modality discrepancy caused by different types of imaging systems. To reduce both intra- and cross_x0002_modality discrepancies, we propose a **Hierarchical Cross_x0002_Modality Disentanglement (Hi-CMD)** method, which automatically disentangles ID-discriminative factors and ID_x0002_excluded factors from visible-thermal images. We only use ID-discriminative factors for robust cross-modality matching without ID-excluded factors such as pose or illumination. To implement our approach, we introduce an ID_x0002_preserving person image generation network and a hierarchical feature learning module. Our generation network learns the disentangled representation by generating a new cross-modality image with different poses and illuminations while preserving a person's identity. At the same time, the feature learning module enables our model to explicitly extract the common ID-discriminative characteristic be_x0002_tween visible-infrared images.

CVPR 2020. Seokeun Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim.

- ReID #2

Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification

- Results



Datasets	RegDB [21]			SYSU-MM01 [33]		
Methods	R=1	R=10	mAP	R=1	R=10	mAP
HOG [3]	13.49	33.22	10.31	2.76	18.25	4.24
LOMO [15]	0.85	2.47	2.28	1.75	14.14	3.48
MLBP [16]	2.02	7.33	6.77	2.12	16.23	3.86
GSM [17]	17.28	34.47	15.06	5.29	33.71	8.00
SVDNet [27]	17.24	34.12	19.04	14.64	53.28	15.17
PCB [28]	18.32	36.42	20.13	16.43	54.06	16.26
One stream [33]	13.11	32.98	14.02	12.04	49.68	13.67
Two stream [33]	12.43	30.36	13.42	11.65	47.99	12.85
Zero padding [33]	17.75	34.21	18.90	14.80	54.12	15.95
TONE [37]	16.87	34.03	14.92	12.52	50.72	14.42
TONE+HCML[37]	24.44	47.53	20.80	14.32	53.16	16.16
BCTR [39]	32.67	57.64	30.99	16.12	54.90	19.15
BDTR [39]	33.47	58.42	31.83	17.01	55.43	19.66
eBDTR(alex) [38]	34.62	58.96	33.46	22.42	64.61	24.11
eBDTR(resnet) [38]	31.83	56.12	33.18	27.82	67.34	28.42
cmGAN [2]	-	-	-	26.97	67.51	27.80
D2RL [31]	43.40	66.10	44.10	**28.90**	**70.60**	**29.20**
HSME [9]	41.34	65.21	38.82	18.03	58.31	19.98
D-HSME [9]	**50.85**	**73.36**	**47.00**	20.68	62.74	23.12
Ours (Hi-CMD)	**70.93**	**86.39**	**66.04**	**34.94**	**77.58**	**35.94**

Table 1. Comparison with the state-of-the-arts on RegDB and SYSU-MM01 datasets. Re-identification rates (%) at rank R and mAP (%). 1^{st} and 2^{nd} best results are indicated by **red** and **blue** color, respectively.

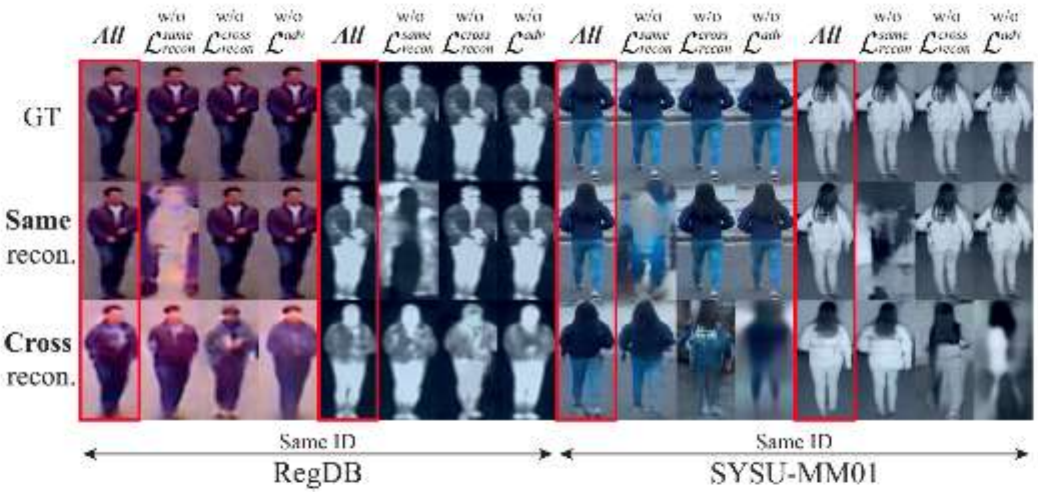
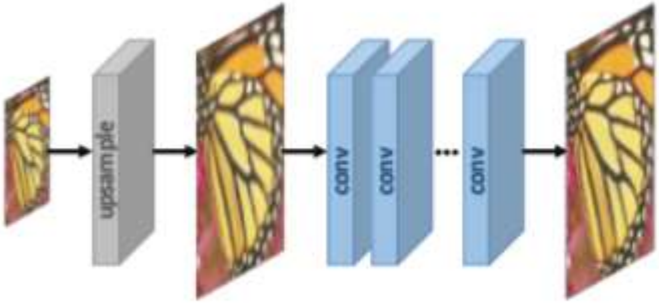

All
w/o $\mathcal{L}^{same}_{recon}$
w/o $\mathcal{L}^{cross}_{recon}$
w/o $\mathcal{L}^{adv}$
All
w/o $\mathcal{L}^{same}_{recon}$
w/o $\mathcal{L}^{cross}_{recon}$
w/o $\mathcal{L}^{adv}$
All
w/o $\mathcal{L}^{same}_{recon}$
w/o $\mathcal{L}^{cross}_{recon}$
w/o $\mathcal{L}^{adv}$
All
w/o $\mathcal{L}^{same}_{recon}$
w/o $\mathcal{L}^{cross}_{recon}$
w/o $\mathcal{L}^{adv}$
GT
Same recon.
Cross recon.
Same ID
RegDB
Same ID
SYSU-MM01

Figure 4. Qualitative comparison between image generation networks for various loss combinations on RegDB and SYSU-MM01. Zoom in for best view.

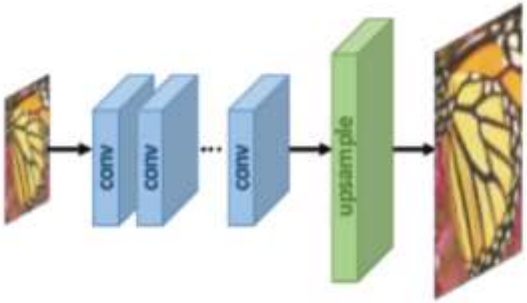
超分辨率

超分辨率：由一幅低分辨率图像或图像序列恢复出高分辨率图像	
常用数据集	Set5、BSD100、Urban100、Manga109、DIV2K2017、CelebA等
常用指标	PSNR，测量有损变换的重建质量
	SSIM，测量图像之间的结构相似性
	MOS，平均意见得分
	IFC ，信息保真标准
代表算法	1.Xiaoou Tang(港中文)——SRCNN、FSRCNN、ESRGAN、EDVR、DNI
	2.Kyoung Mu Lee(首尔国立大学CV Lab)——DRCN、VDSR
	3.Ming-Hsuan Yang(UC Merced)——LapSRN
	4.Thomas Huang(UIUC)——WDSR
	5.Twitter——SRGAN

- Super-resolution Frameworks based on the employed **upsampling operations** and their locations in the model



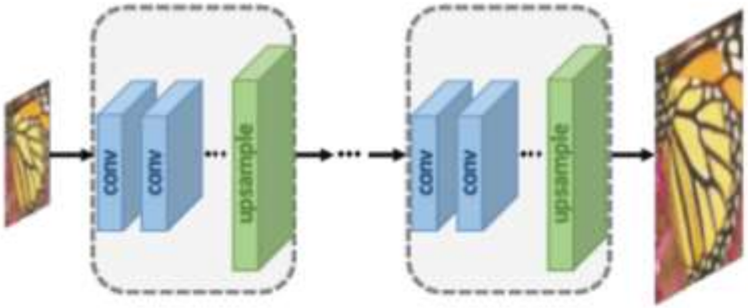
(a) Pre-upsampling SR



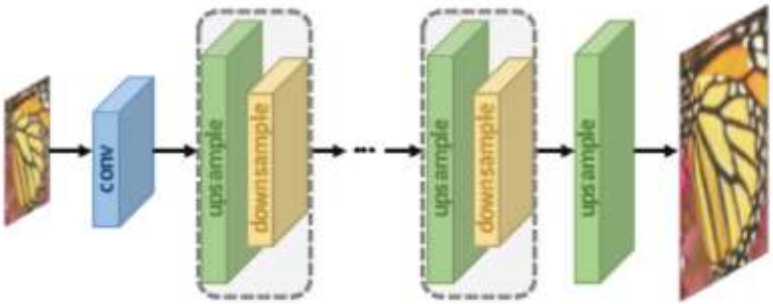
(b) Post-upsampling SR

前端升采样网络

后端升采样网络



(c) Progressive upsampling SR



(d) Iterative up-and-down Sampling SR

渐进式升采样网络

升降采样迭代式超分网络

TPAMI 2020. Deep Learning for Image Super-resolution: A Survey

- **可学习的升采样方法**
  1. 转置卷积（Transposed Convolution），反卷积（Deconvolution）
  2. 亚像素（Sub-pixel）卷积

- **全局和局部网络结构设计**
  1. 残差学习 （Residual Learning）
  2. 递归学习 （Recursive Learning）
  3. 多支路学习 （Multi-path Learning）
  4. 稠密连接 （Dense Connections）
  5. 通道重缩放 （Channel Attention）
  6. 高级卷积结构 （Advanced Convolution）
  7. 像素递归学习 （Pixel Recursive Learning）
  8. 金字塔池化 （Pyramid Pooling）
  9. 小波域变换 （Wavelet Transformation）

- **损失函数设计**
  1. 像素级 （Pixel Loss）
  2. 内容损失 （Content Loss）
  3. 纹理损失 （Texture Loss）
  4. 对抗生成损失 （ Adversarial Loss）
  5. 往复一致性保持损失 （Cycle Consistency Loss）
  6. 全变分损失 （Total Variation Loss）
  7. 基于先验知识的损失 （Prior-based Loss）

- State-of-the-art Super-resolution Models

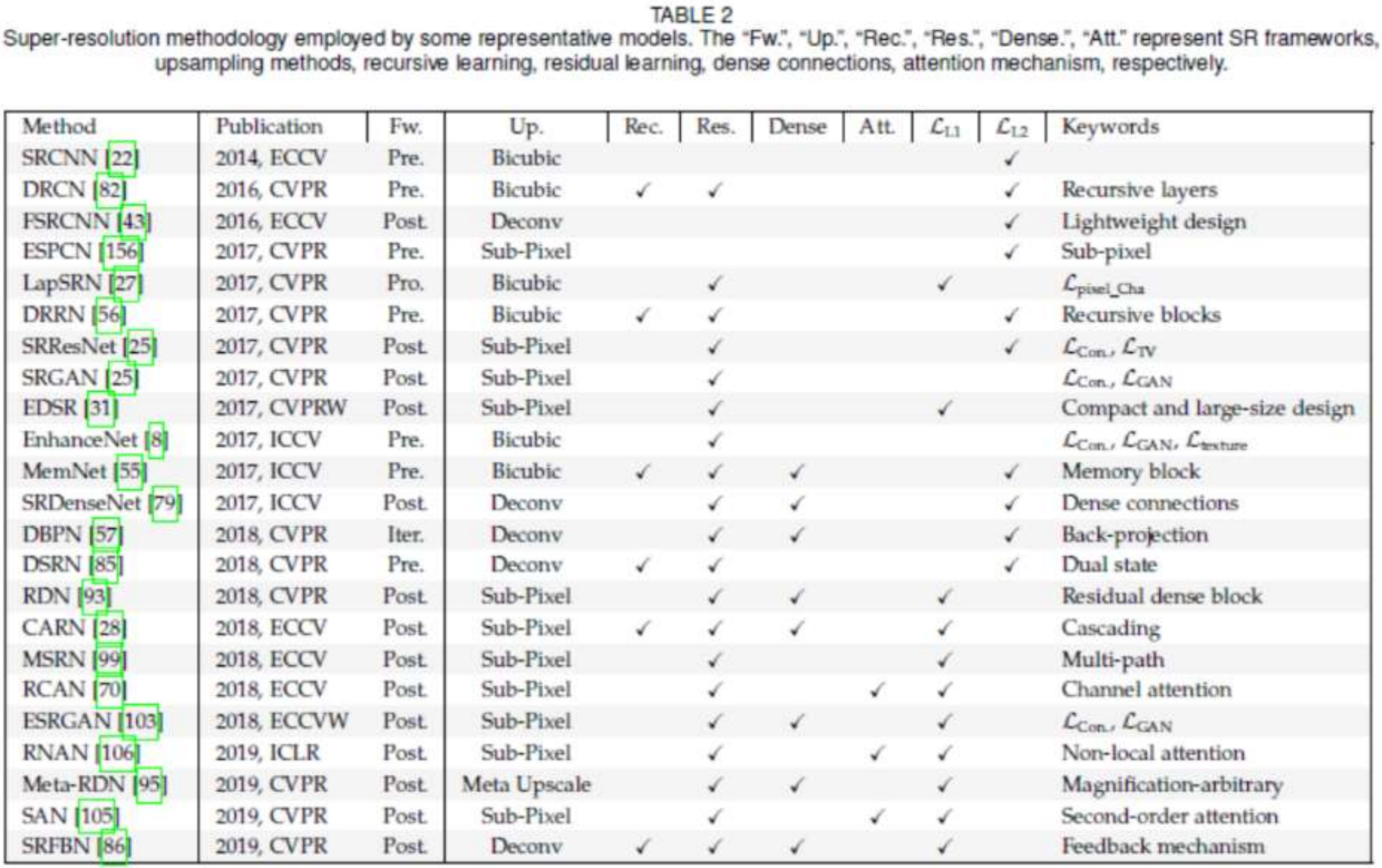


TABLE 2
Super-resolution methodology employed by some representative models. The "Fw.", "Up.", "Rec.", "Res.", "Dense.", "Att." represent SR frameworks, upsampling methods, recursive learning, residual learning, dense connections, attention mechanism, respectively.

Method	Publication	Fw.	Up.	Rec.	Res.	Dense	Att.	$\mathcal{L}_{L1}$	$\mathcal{L}_{L2}$	Keywords
SRCNN [22]	2014, ECCV	Pre.	Bicubic						✓	
DRCN [82]	2016, CVPR	Pre.	Bicubic	✓	✓				✓	Recursive layers
FSRCNN [43]	2016, ECCV	Post.	Deconv						✓	Lightweight design
ESPCN [156]	2017, CVPR	Pre.	Sub-Pixel						✓	Sub-pixel
LapSRN [27]	2017, CVPR	Pro.	Bicubic		✓			✓		$\mathcal{L}_{pixel\_Cha}$
DRRN [56]	2017, CVPR	Pre.	Bicubic	✓	✓				✓	Recursive blocks
SRResNet [25]	2017, CVPR	Post.	Sub-Pixel		✓				✓	$\mathcal{L}_{Con.}$, $\mathcal{L}_{TV}$
SRGAN [25]	2017, CVPR	Post.	Sub-Pixel		✓					$\mathcal{L}_{Con.}$, $\mathcal{L}_{GAN}$
EDSR [31]	2017, CVPRW	Post.	Sub-Pixel		✓			✓		Compact and large-size design
EnhanceNet [8]	2017, ICCV	Pre.	Bicubic		✓					$\mathcal{L}_{Con.}$, $\mathcal{L}_{GAN}$, $\mathcal{L}_{texture}$
MemNet [55]	2017, ICCV	Pre.	Bicubic	✓	✓	✓			✓	Memory block
SRDenseNet [79]	2017, ICCV	Post.	Deconv		✓	✓			✓	Dense connections
DBPN [57]	2018, CVPR	Iter.	Deconv		✓	✓			✓	Back-projection
DSRN [85]	2018, CVPR	Pre.	Deconv	✓	✓				✓	Dual state
RDN [93]	2018, CVPR	Post.	Sub-Pixel		✓	✓		✓		Residual dense block
CARN [28]	2018, ECCV	Post.	Sub-Pixel	✓	✓	✓		✓		Cascading
MSRN [99]	2018, ECCV	Post.	Sub-Pixel		✓			✓		Multi-path
RCAN [70]	2018, ECCV	Post.	Sub-Pixel		✓		✓	✓		Channel attention
ESRGAN [103]	2018, ECCVW	Post.	Sub-Pixel		✓	✓		✓		$\mathcal{L}_{Con.}$, $\mathcal{L}_{GAN}$
RNAN [106]	2019, ICLR	Post.	Sub-Pixel		✓		✓	✓		Non-local attention
Meta-RDN [95]	2019, CVPR	Post.	Meta Upscale		✓	✓		✓		Magnification-arbitrary
SAN [105]	2019, CVPR	Post.	Sub-Pixel		✓		✓	✓		Second-order attention
SRFBN [86]	2019, CVPR	Post.	Deconv	✓	✓	✓		✓		Feedback mechanism

- State-of-the-art Super-resolution Models

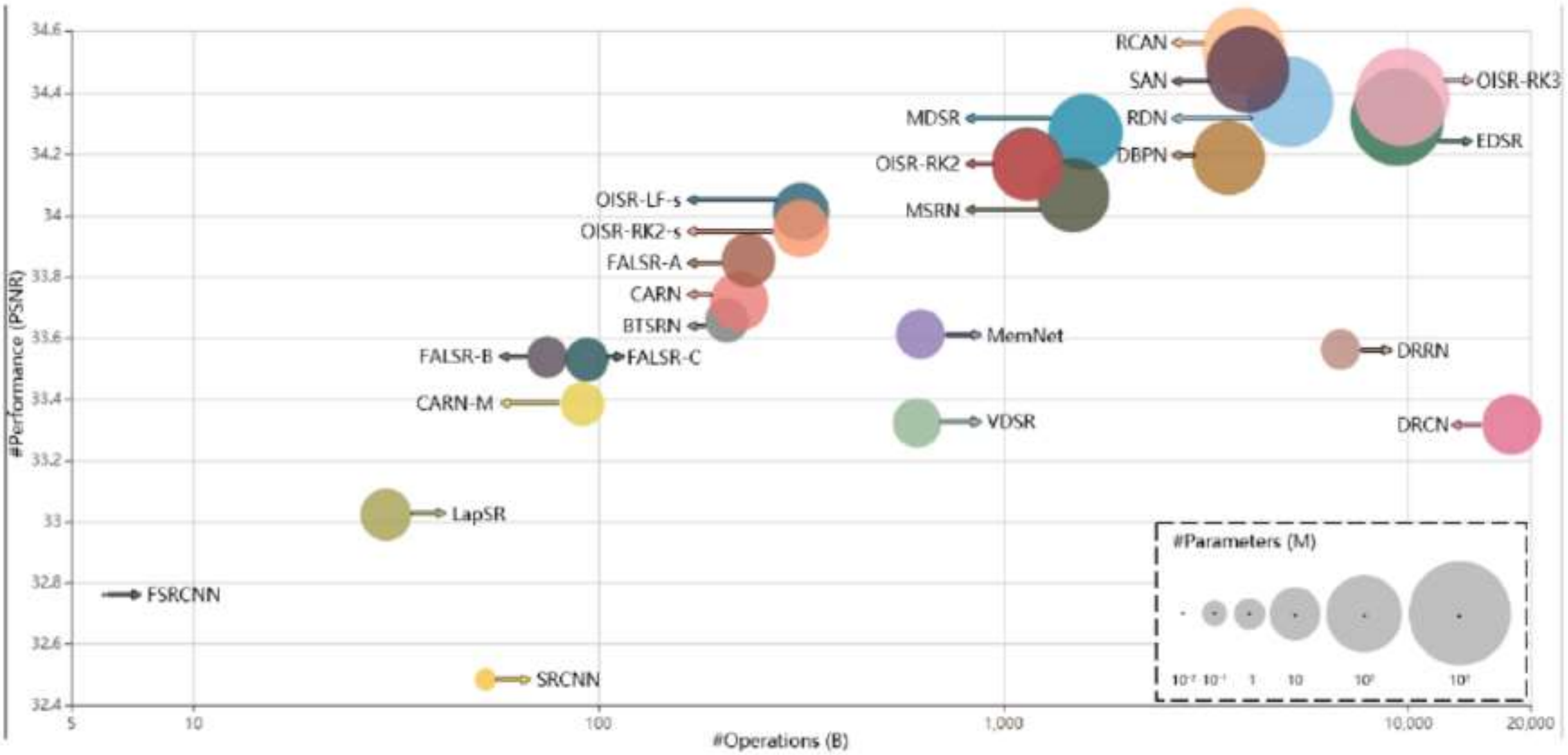


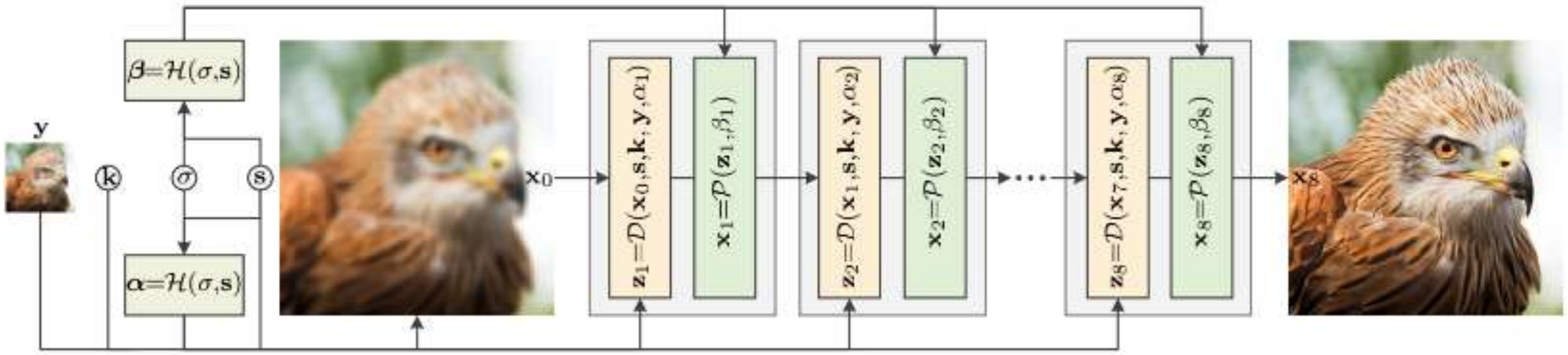
Fig. 8. Super-resolution benchmarking. The x-axis and the y-axis denote the Multi-Adds and PSNR, respectively, and the circle size represents the number of parameters.

- The accuracy is measured by the mean of the PSNR on 4 benchmark datasets (i.e., Set5 [48], Set14 [49], B100 [40] and Urban100 [50]).
- And the model size and computational cost are calculated with PyTorch-OpCounter [157], where the output resolution is 720p (i.e., 1080*720).
- All statistics are derived from the original papers or calculated on official models, with a scaling factor of 2.

- 超分辨率 #1

Deep Unfolding Network for Image Super-Resolution

- Abstract & Overall Architecrure


$\mathbf{y}$
$\beta=\mathcal{H}(\sigma,\mathbf{s})$
$\alpha=\mathcal{H}(\sigma,\mathbf{s})$
$\mathbf{k}$
σ
$\mathbf{s}$
$\mathbf{x}_0$
$\mathbf{z}_1=\mathcal{D}(\mathbf{x}_0,\mathbf{s},\mathbf{k},\mathbf{y},\alpha_1)$
$\mathbf{x}_1=\mathcal{P}(\mathbf{z}_1,\beta_1)$
$\mathbf{z}_2=\mathcal{D}(\mathbf{x}_1,\mathbf{s},\mathbf{k},\mathbf{y},\alpha_2)$
$\mathbf{x}_2=\mathcal{P}(\mathbf{z}_2,\beta_2)$
$\cdots$
$\mathbf{z}_8=\mathcal{D}(\mathbf{x}_7,\mathbf{s},\mathbf{k},\mathbf{y},\alpha_8)$
$\mathbf{x}_8=\mathcal{P}(\mathbf{z}_8,\beta_8)$
$\mathbf{x}_8$

Learning-based single image super-resolution (SISR) methods are continuously showing superior effectiveness and efficiency over traditional model-based methods, largely due to the end-to-end training. However, different from model-based methods that can handle the SISR problem with different scale factors, blur kernels and noise levels under a unified MAP (maximum a posteriori) framework, learning-based methods generally lack such flexibility. To address this issue, this paper proposes an end-to-end trainable unfolding network which leverages both learning-based methods and model-based methods. Specifically, by unfolding the MAP inference via a half-quadratic splitting algorithm, a fixed number of iterations consisting of alternately solving a data subproblem and a prior subproblem can be obtained. The two subproblems then can be solved with neural modules, resulting in an end-to-end trainable, iterative network. As a result, the proposed network inherits the flexibility of model-based methods to super-resolve blurry, noisy images for different scale factors via a single model, while maintaining the advantages of learning-based methods.

CVPR 2020. Kai Zhang, Luc Van Gool, and Radu Timofte.

- 超分辨率 #1

Deep Unfolding Network for Image Super-Resolution

- Results

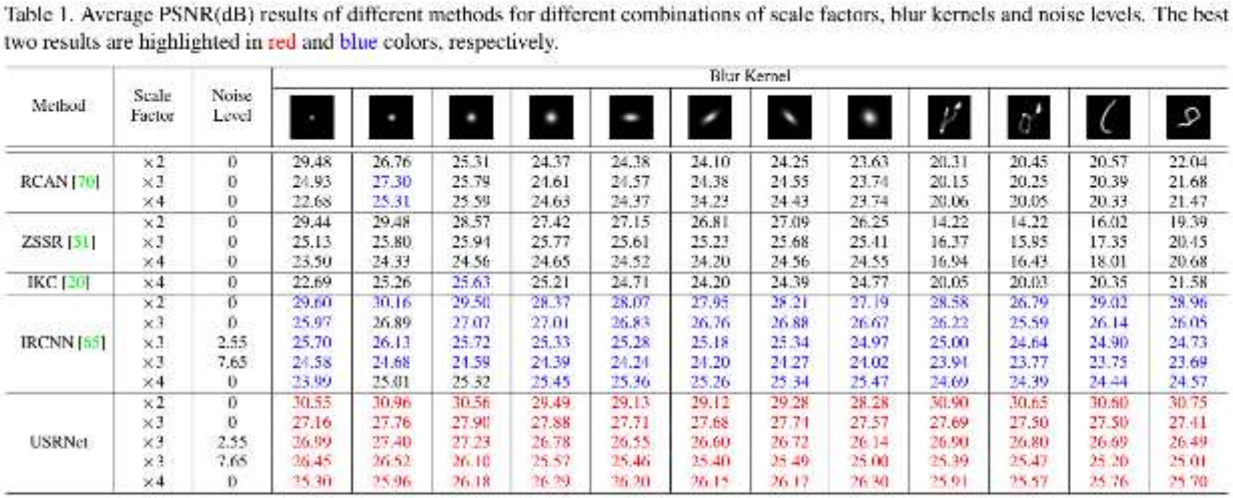
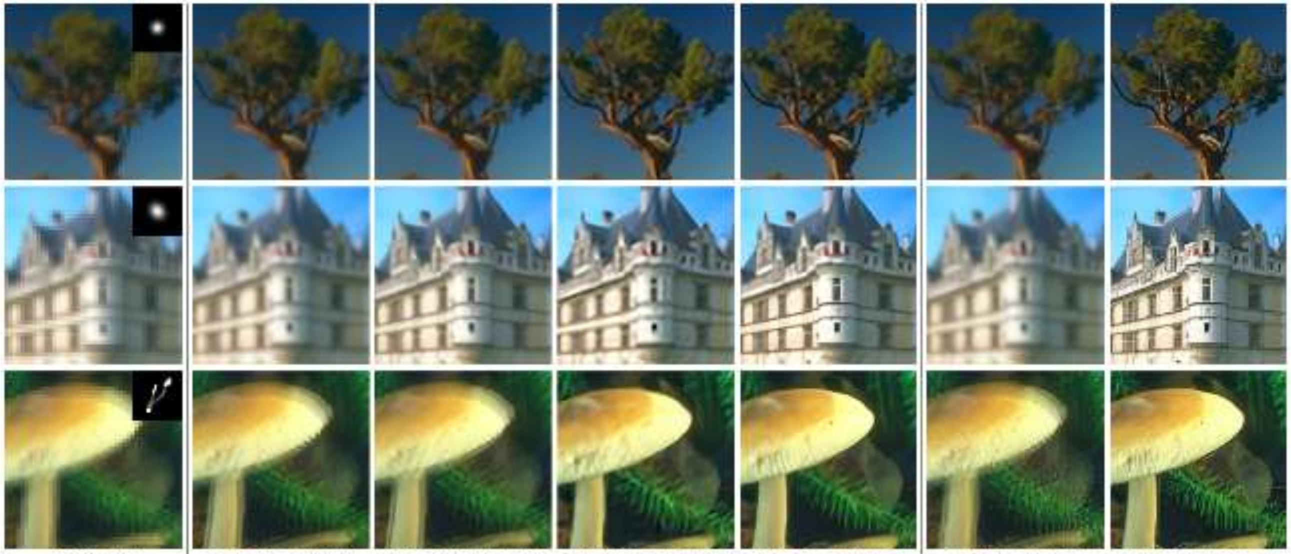


Table 1. Average PSNR(dB) results of different methods for different combinations of scale factors, blur kernels and noise levels. The best two results are highlighted in red and blue colors, respectively.

Method	Scale Factor	Noise Level	Blur Kernel											
RCAN [70]	×2	0	29.48	26.76	25.31	24.37	24.38	24.10	24.25	23.63	20.31	20.45	20.57	22.04
	×3	0	24.93	27.30	25.79	24.61	24.57	24.38	24.55	23.74	20.15	20.25	20.39	21.68
	×4	0	22.68	25.31	25.59	24.63	24.37	24.23	24.43	23.74	20.06	20.05	20.33	21.47
ZSSR [51]	×2	0	29.44	29.48	28.57	27.42	27.15	26.81	27.09	26.25	14.22	14.22	16.02	19.39
	×3	0	25.13	25.80	25.94	25.77	25.61	25.23	25.68	25.41	16.37	15.95	17.35	20.45
	×4	0	23.50	24.33	24.56	24.65	24.52	24.20	24.56	24.55	16.94	16.43	18.01	20.68
IKC [20]	×4	0	22.69	25.26	25.63	25.21	24.71	24.20	24.39	24.77	20.05	20.03	20.35	21.58
IRCNN [65]	×2	0	29.60	30.16	29.50	28.37	28.07	27.95	28.21	27.19	28.58	26.79	29.02	28.96
	×3	0	25.97	26.89	27.07	27.01	26.83	26.76	26.88	26.67	26.22	25.59	26.14	26.05
	×3	2.55	25.70	26.13	25.72	25.33	25.28	25.18	25.34	24.97	25.00	24.64	24.90	24.73
	×3	7.65	24.58	24.68	24.59	24.39	24.24	24.20	24.27	24.02	23.94	23.77	23.75	23.69
	×4	0	23.99	25.01	25.32	25.45	25.36	25.26	25.34	25.47	24.69	24.39	24.44	24.57
USRNet	×2	0	30.55	30.96	30.56	29.49	29.13	29.12	29.28	28.28	30.90	30.65	30.60	30.75
	×3	0	27.16	27.76	27.90	27.88	27.71	27.68	27.71	27.57	27.69	27.50	27.50	27.41
	×3	2.55	26.99	27.40	27.23	26.78	26.55	26.60	26.72	26.14	26.90	26.80	26.69	26.49
	×3	7.65	26.45	26.52	26.10	25.57	25.46	25.40	25.49	25.00	25.39	25.47	25.20	25.01
	×4	0	25.30	25.96	26.18	26.29	26.20	26.15	26.17	26.30	25.91	25.57	25.76	25.70



PSNR(dB) Zoomed LR (×4)

24.92/22.53/21.88 RCAN [70]

25.24/23.75/21.92 IKC [20]

25.42/23.55/26.45 IRCNN [65]

25.82/24.80/27.39 **USRNet** (ours)

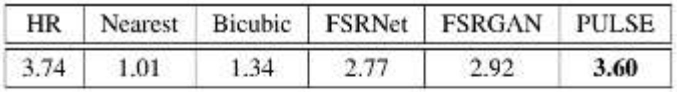
24.80/22.44/21.51 RankSRGAN [69]

24.26/23.25/25.00 **USRGAN** (ours)

- 超分辨率 #2

PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models

- Abstract & Results



HR	Nearest	Bicubic	FSRNet	FSRGAN	PULSE
3.74	1.01	1.34	2.77	2.92	**3.60**

Table 1. MOS Score for various algorithms at 128 × 128. Higher is better.

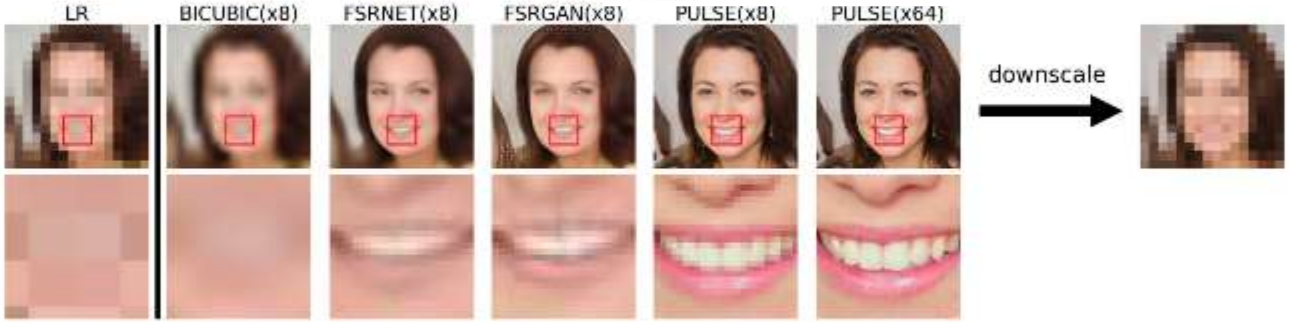


Figure 5. Comparison of PULSE with bicubic upscaling, FSRNet, and FSRGAN. In the first image, PULSE adds a messy patch in the hair to match the two dark diagonal pixels visible in the middle of the zoomed in LR image.

The primary aim of single-image super-resolution is to construct a HR image from a corresponding LR input. In previous approaches, which have generally been supervised, the training objective typically measures a pixel-wise average distance between the super-resolved (SR) and HR images. Optimizing such metrics often leads to blurring, especially in high variance (detailed) regions. We propose an alternative formulation of the super-resolution problem based on creating realistic SR images that downscale correctly. We present a novel super-resolution algorithm addressing this problem, **PULSE (Photo Upsampling via Latent Space Exploration)**. It accomplishes this in an entirely self-supervised fashion and is not confined to a specific degradation operator used during training, unlike previous methods (which require training on databases of LR-HR image pairs for supervised learning). Instead of starting with the LR image and slowly adding detail, PULSE traverses the high-resolution natural image manifold, searching for images that downscale to the original LR image. This is formalized through the “downscaling loss,” which guides exploration through the latent space of a generative model. By leveraging properties of high-dimensional Gaussians, we restrict the search space to guarantee that our outputs are realistic.

CVPR 2020. Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, Cynthia Rudin.

- 超分辨率 #3

Meta-SR: A Magnification-Arbitrary Network for Super-Resolution

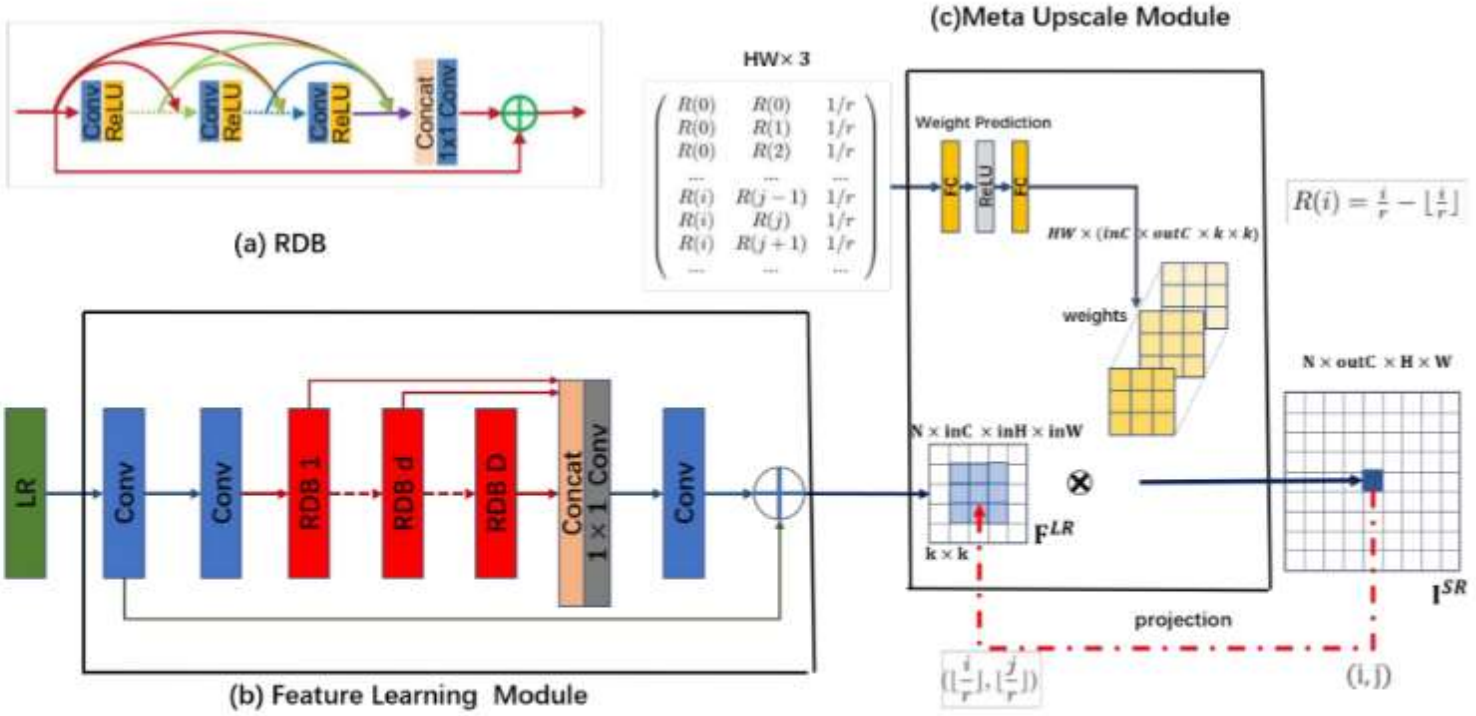
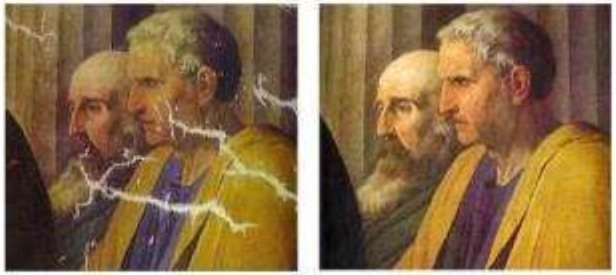


Figure 1. An instance of our Meta-SR based on RDN [36]. We also call the network Meta-RDN. (a) The Residual Dense Block proposed by RDN [36]. (b) The Feature Learning Module which generates the shared feature maps for arbitrary scale factor. (c) For each pixel on the SR image, we project it onto the LR image. The proposed Meta-Upscale Module takes a sequence of coordinate-related and scale-related vectors as input to predict the weights for convolution filters. By doing the convolution operation, our Meta-Upscale finally generate the HR image.

CVPR 2020. Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, Jian Sun

Inpainting

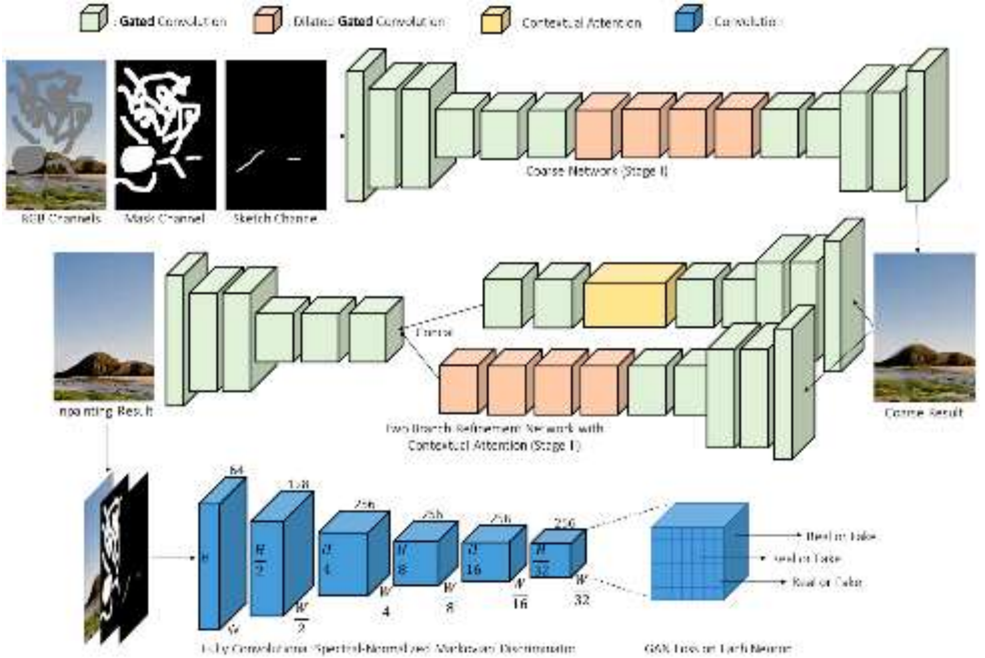


Inpainting：对受到损坏的图像进行修复重建或者去除图像中的多余物体	
常用数据集	Places、CelebA、Paris StreetView、ImageNet等
常用指标	PSNR，测量重建图像的质量
	SSIM，测量图像之间的结构相似性
	L1/L2，修复结果与原图的差异
代表算法	1.Thomas Huang, Jiahui Yu(UIUC)——GatedConv、ContextAttention
	2.Alexei A. Efros(UC Berkeley)——Context Encoder
	3.Adobe——MNPS
	4.Satoshi Iizuka(Waseda University)——GLCIC
	5.Guilin Liu(NVIDIA)——PartialConv
	6.Wangmeng Zuo(哈工大)——Shift-Net

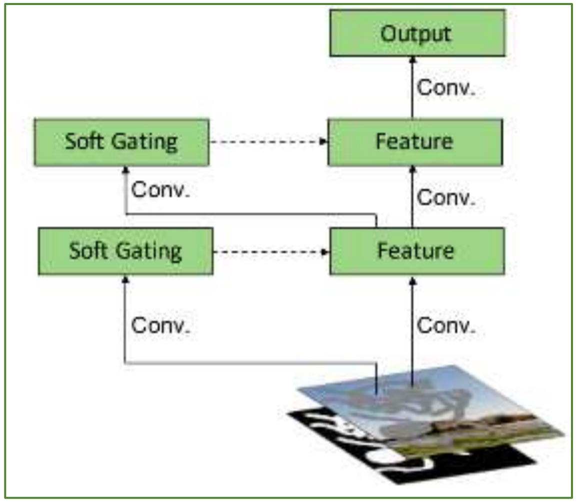
- Inpainting #1

Free-Form Image Inpainting with Gated Convolution

- Abstract & Architecture



Gated Convolution



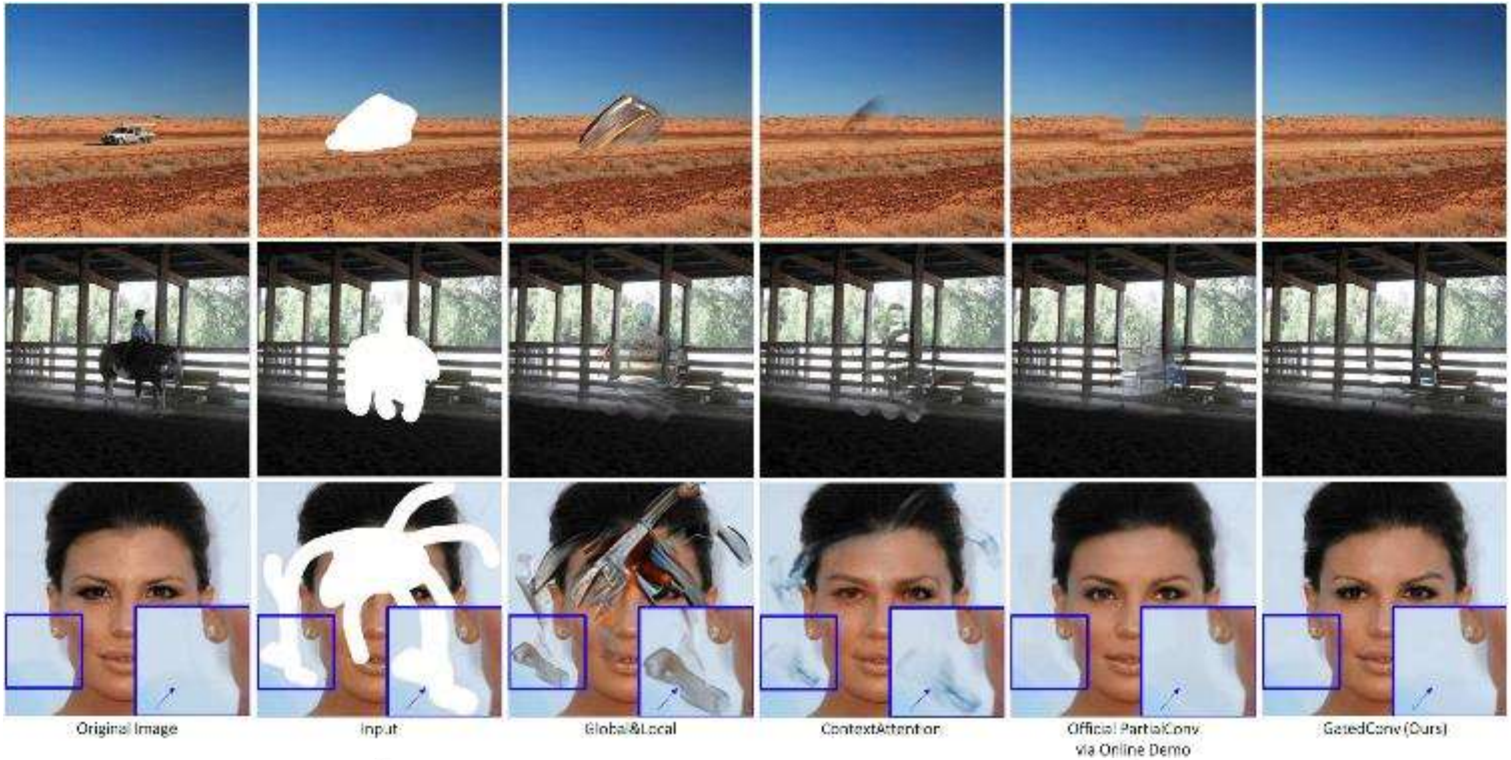
This paper presents a generative image inpainting system to complete images with **free-form mask** and **guidance**. The system is based on **gated convolutions** learned from millions of images without additional labelling efforts. The proposed gated convolution solves the issue of vanilla convolution that treats all input pixels as valid ones, generalizes partial convolution by providing a learnable dynamic feature selection mechanism for each channel at each spatial location across all layers. Moreover, as free-form masks may appear anywhere in images with any shape, global and local GANs designed for a single rectangular mask are not applicable. Thus, it also presents a patch-based GAN loss, named **SN-PatchGAN**, by applying spectral-normalized discriminator on dense image patches. SN-PatchGAN is simple in formulation, fast and stable in training. This system helps user quickly remove distracting objects, modify image layouts, clear watermarks and edit faces.

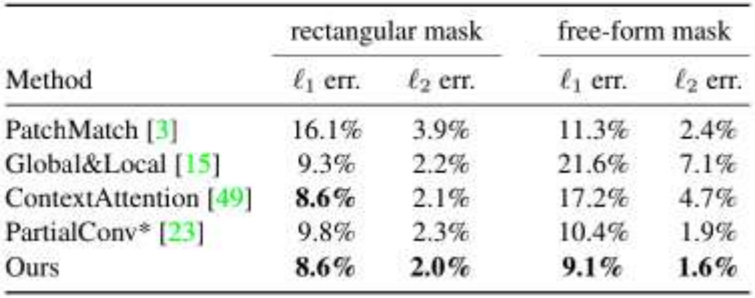
ICCV 2019. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang.

- Inpainting #1

Free-Form Image Inpainting with Gated Convolution

- Results



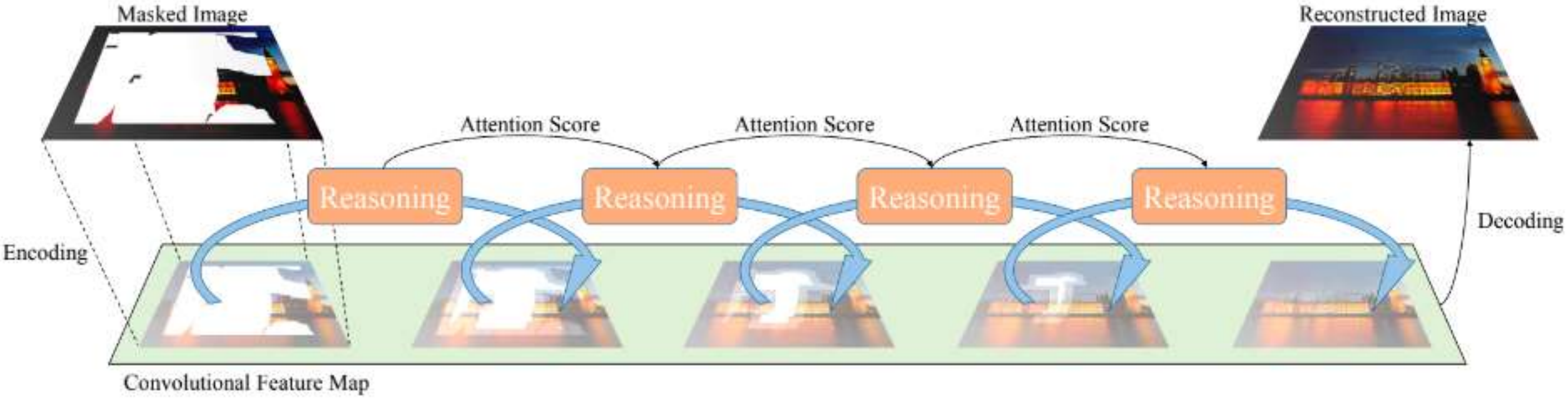


	rectangular mask		free-form mask	
Method	ℓ_1 err.	ℓ_2 err.	ℓ_1 err.	ℓ_2 err.
PatchMatch [3]	16.1%	3.9%	11.3%	2.4%
Global&Local [15]	9.3%	2.2%	21.6%	7.1%
ContextAttention [49]	**8.6%**	2.1%	17.2%	4.7%
PartialConv* [23]	9.8%	2.3%	10.4%	1.9%
Ours	**8.6%**	**2.0%**	**9.1%**	**1.6%**

- Inpainting #2

Recurrent Feature Reasoning for Image Inpainting

- Abstract & Overview



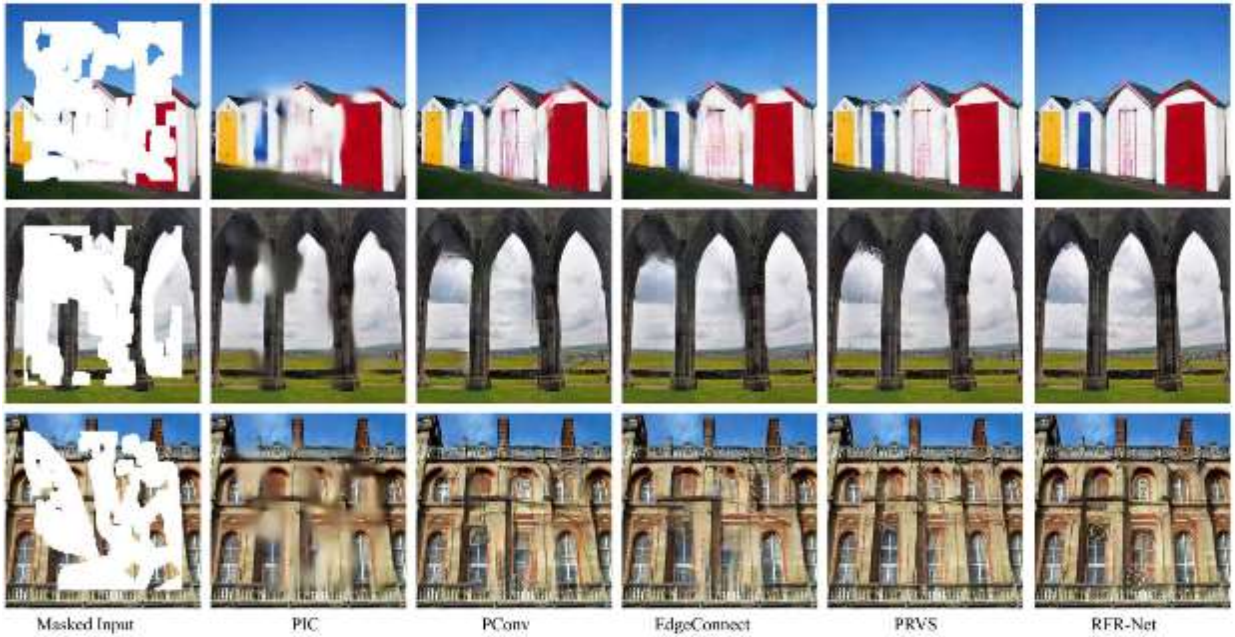
Existing inpainting methods have achieved promising performance for recovering regular or small image defects. However, filling in large continuous holes remains difficult due to the lack of constraints for the hole center. In this paper, we devise a **Recurrent Feature Reasoning (RFR)** network which is mainly constructed by a plug-and-play Recurrent Feature Reasoning module and a **Knowledge Consistent Attention (KCA)** module. Analogous to how humans solve puzzles (i.e., first solve the easier parts and then use the results as additional information to solve difficult parts), the RFR module recurrently infers the hole boundaries of the convolutional feature maps and then uses them as clues for further inference. The module progressively strengthens the constraints for the hole center and the results become explicit. To capture information from distant places in the feature map for RFR, we further develop KCA and incorporate it in RFR.

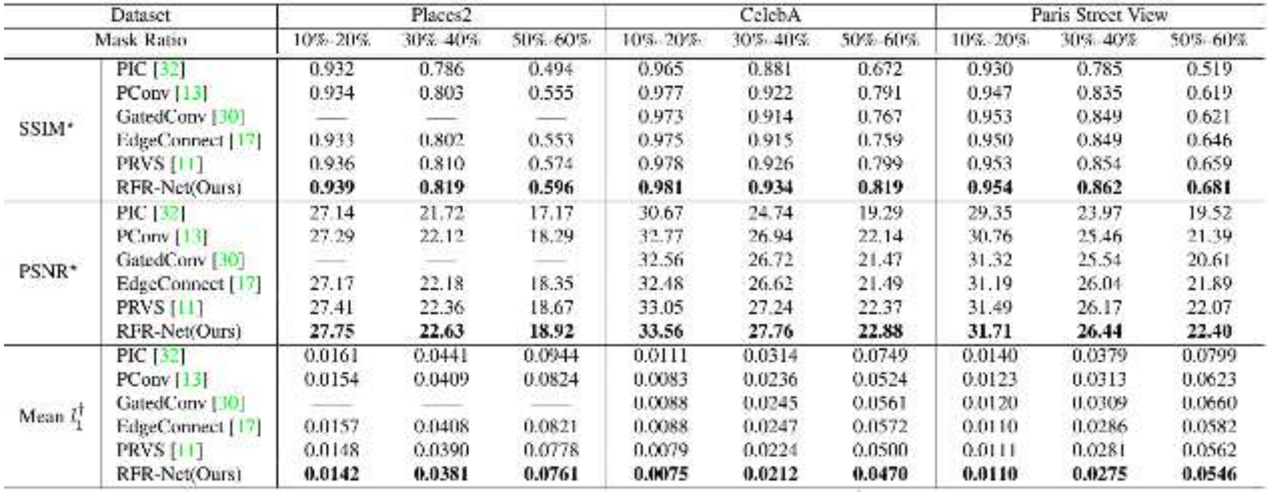
CVPR 2020. Jingyuan Li, Ning Wang, Lefei Zhang*, Bo Du, and Dacheng Tao.

- Inpainting #2

Recurrent Feature Reasoning for Image Inpainting

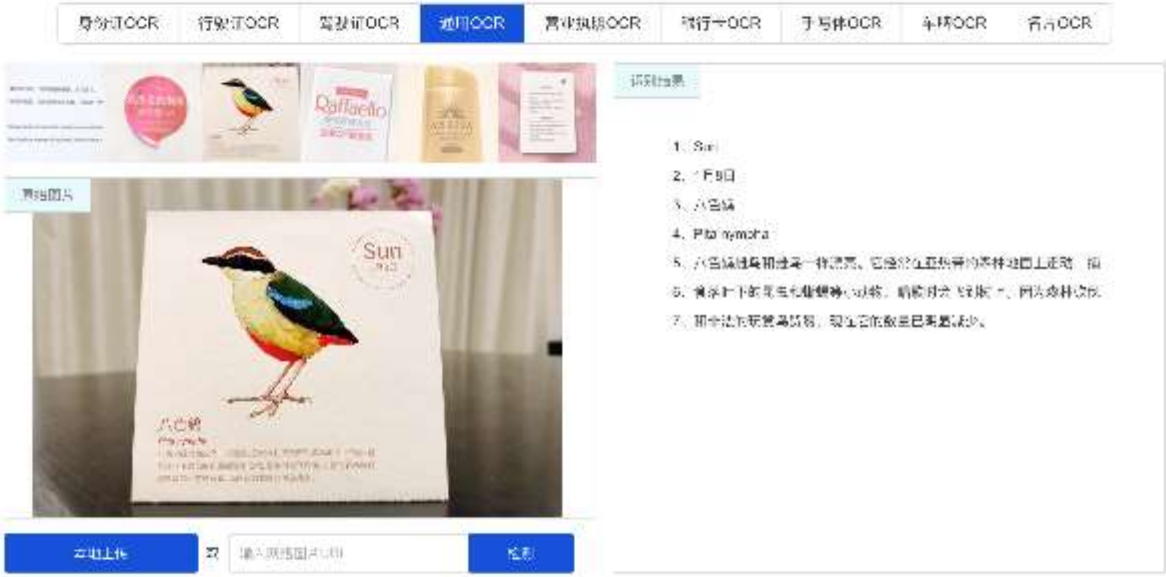
- Results





Dataset		Places2			CelebA			Paris Street View		
Mask Ratio		10%-20%	30%-40%	50%-60%	10%-20%	30%-40%	50%-60%	10%-20%	30%-40%	50%-60%
SSIM*	PIC [32]	0.932	0.786	0.494	0.965	0.881	0.672	0.930	0.785	0.519
	PConv [13]	0.934	0.803	0.555	0.977	0.922	0.791	0.947	0.835	0.619
	GatedConv [30]	——	——	——	0.973	0.914	0.767	0.953	0.849	0.621
	EdgeConnect [17]	0.933	0.802	0.553	0.975	0.915	0.759	0.950	0.849	0.646
	PRVS [11]	0.936	0.810	0.574	0.978	0.926	0.799	0.953	0.854	0.659
	RFR-Net(Ours)	**0.939**	**0.819**	**0.596**	**0.981**	**0.934**	**0.819**	**0.954**	**0.862**	**0.681**
PSNR*	PIC [32]	27.14	21.72	17.17	30.67	24.74	19.29	29.35	23.97	19.52
	PConv [13]	27.29	22.12	18.29	32.77	26.94	22.14	30.76	25.46	21.39
	GatedConv [30]	——	——	——	32.56	26.72	21.47	31.32	25.54	20.61
	EdgeConnect [17]	27.17	22.18	18.35	32.48	26.62	21.49	31.19	26.04	21.89
	PRVS [11]	27.41	22.36	18.67	33.05	27.24	22.37	31.49	26.17	22.07
	RFR-Net(Ours)	**27.75**	**22.63**	**18.92**	**33.56**	**27.76**	**22.88**	**31.71**	**26.44**	**22.40**
Mean $l_1^\dagger$	PIC [32]	0.0161	0.0441	0.0944	0.0111	0.0314	0.0749	0.0140	0.0379	0.0799
	PConv [13]	0.0154	0.0409	0.0824	0.0083	0.0236	0.0524	0.0123	0.0313	0.0623
	GatedConv [30]	——	——	——	0.0088	0.0245	0.0561	0.0120	0.0309	0.0660
	EdgeConnect [17]	0.0157	0.0408	0.0821	0.0088	0.0247	0.0572	0.0110	0.0286	0.0582
	PRVS [11]	0.0148	0.0390	0.0778	0.0079	0.0224	0.0500	0.0111	0.0281	0.0562
	RFR-Net(Ours)	**0.0142**	**0.0381**	**0.0761**	**0.0075**	**0.0212**	**0.0470**	**0.0110**	**0.0275**	**0.0546**

OCR



1、Sun

2、7月9日

3、八色鸫

4、Pitta nympha

5、八色鸫性格相当羞一样谨慎，它经常在亚热带的森林地面上走动，捕

6、食落叶下的昆虫和蠕虫等小动物，[illegible]，因为森林砍伐

7、和非法的捕捉等原因，现在它的数量已经严重减少。

OCR：识别图片中的文字

常用数据集	MNIST、CTW、ICDAR、MSRA-TD500、SVT、COCO-Text等
常用指标	Recall，Precision，F-score
代表算法	1.Cong Yao, Xiang Bai(华科)——CRNN、EAST、ASTER、TextBoxes++
	2.Shuigeng Zhou(复旦大学)——FAN、AON、EP
	3.Kwan-Yee K. Wong(香港大学)——Char-Net、STAR-Net、SAFE
	4.Chunhua Shen(阿德莱德大学)——SAR、TextSR
	5.Bo Xu(中国科学院自动化研究所)——NRTR

- OCR #1

Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning

- Abstract & Architecrure

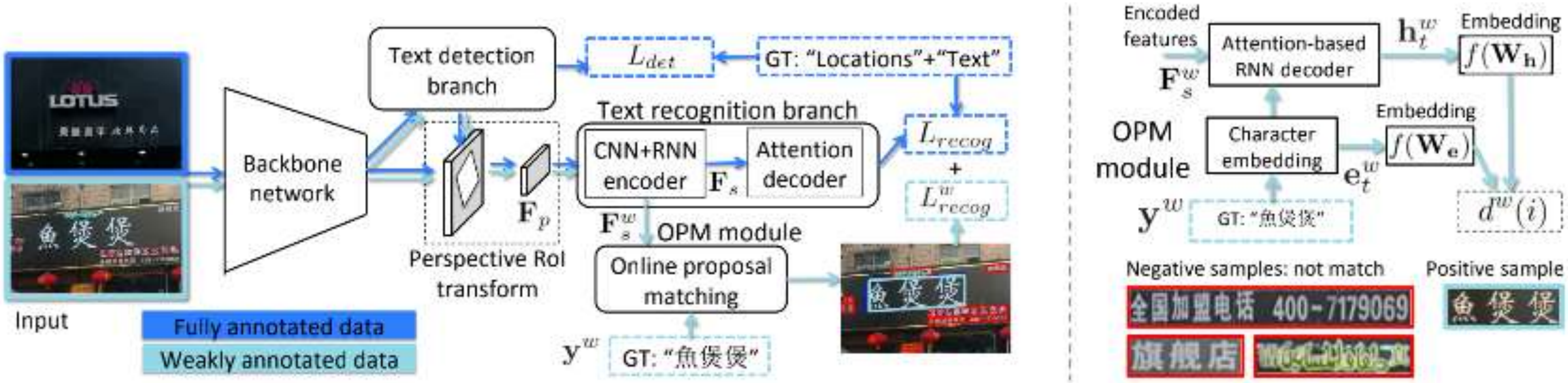


Figure 4: (Left) the overall architecture of the proposed partially supervised end-to-end text reading model. (Right) online proposal matching.

Most existing text reading benchmarks make it difficult to evaluate the performance of more advanced deep learning models in large vocabularies due to the limited amount of training data. To address this issue, we introduce a new large-scale text reading benchmark dataset named **Chinese Street View Text (C-SVT)** with 430, 000 street view images, which is at least 14 times as large as the existing Chinese text reading benchmarks. To recognize Chinese text in the wild while keeping large-scale datasets labeling cost-effective, we propose to annotate one part of the C-SVT dataset (30,000 images) in locations and text labels as full annotations and add 400, 000 more images, where only the corresponding text-of-interest in the regions is given as weak annotations. To exploit the rich information from the weakly annotated data, we design a text reading network in a partially supervised learning framework, which enables to localize and recognize text, learn from fully and weakly annotated data simultaneously. To localize the best matched text proposals from weakly labeled images, we propose an online proposal matching module incorporated in the whole model, spotting the keyword regions by sharing parameters for end-to-end training.

ICCV 2019. Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu.

- OCR #1

Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning

- Results

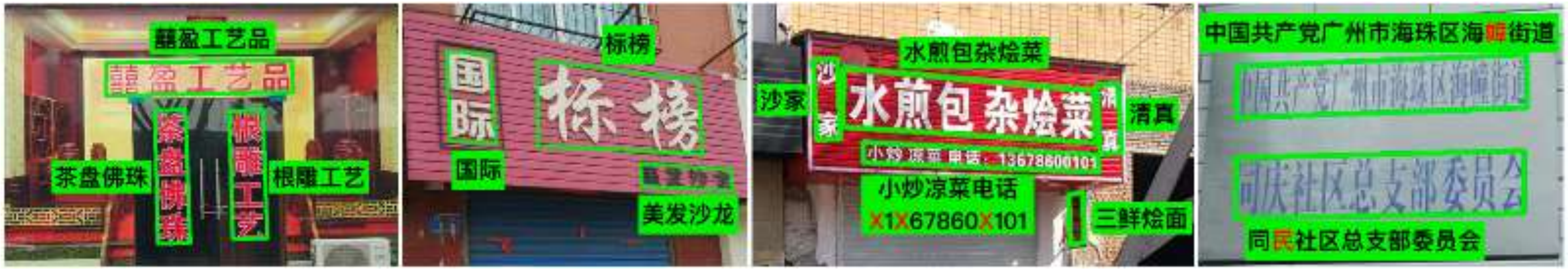
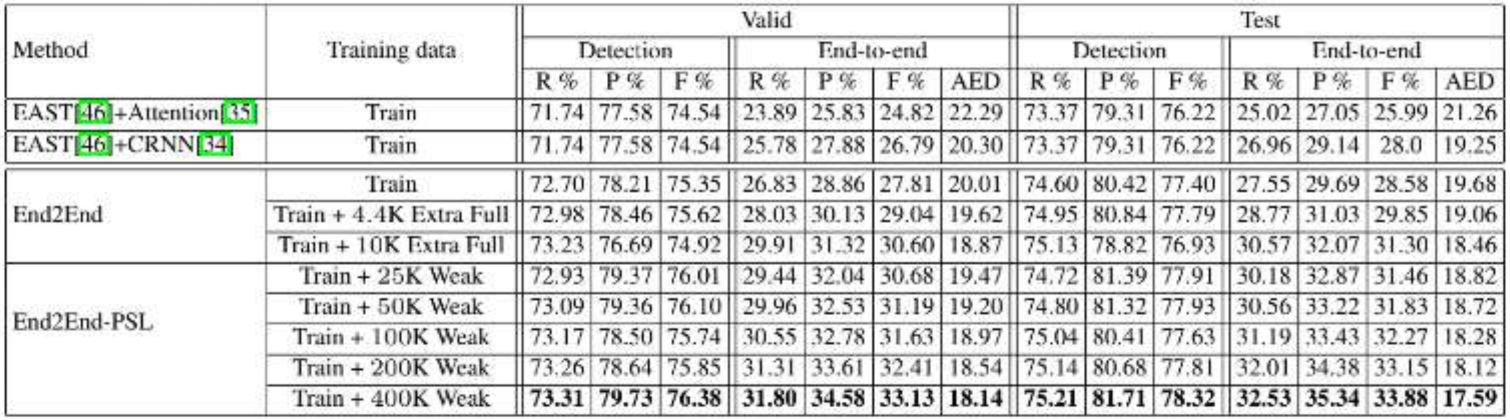
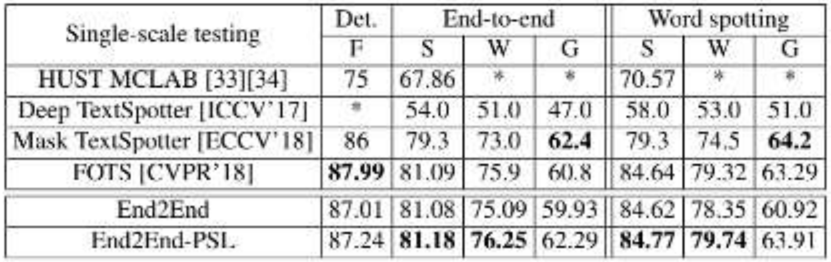


Table 3: The performance of the end-to-end Chinese text reading models on C-SVT . 'PSL' denotes the proposed partially supervised learning algorithm.



Method	Training data	Valid							Test						
		Detection			End-to-end				Detection			End-to-end			
		R %	P %	F %	R %	P %	F %	AED	R %	P %	F %	R %	P %	F %	AED
EAST[46]+Attention[35]	Train	71.74	77.58	74.54	23.89	25.83	24.82	22.29	73.37	79.31	76.22	25.02	27.05	25.99	21.26
EAST[46]+CRNN[34]	Train	71.74	77.58	74.54	25.78	27.88	26.79	20.30	73.37	79.31	76.22	26.96	29.14	28.0	19.25
End2End	Train	72.70	78.21	75.35	26.83	28.86	27.81	20.01	74.60	80.42	77.40	27.55	29.69	28.58	19.68
	Train + 4.4K Extra Full	72.98	78.46	75.62	28.03	30.13	29.04	19.62	74.95	80.84	77.79	28.77	31.03	29.85	19.06
	Train + 10K Extra Full	73.23	76.69	74.92	29.91	31.32	30.60	18.87	75.13	78.82	76.93	30.57	32.07	31.30	18.46
End2End-PSL	Train + 25K Weak	72.93	79.37	76.01	29.44	32.04	30.68	19.47	74.72	81.39	77.91	30.18	32.87	31.46	18.82
	Train + 50K Weak	73.09	79.36	76.10	29.96	32.53	31.19	19.20	74.80	81.32	77.93	30.56	33.22	31.83	18.72
	Train + 100K Weak	73.17	78.50	75.74	30.55	32.78	31.63	18.97	75.04	80.41	77.63	31.19	33.43	32.27	18.28
	Train + 200K Weak	73.26	78.64	75.85	31.31	33.61	32.41	18.54	75.14	80.68	77.81	32.01	34.38	33.15	18.12
	Train + 400K Weak	**73.31**	**79.73**	**76.38**	**31.80**	**34.58**	**33.13**	**18.14**	**75.21**	**81.71**	**78.32**	**32.53**	**35.34**	**33.88**	**17.59**

Table 7: End-to-end results on ICDAR 2015.



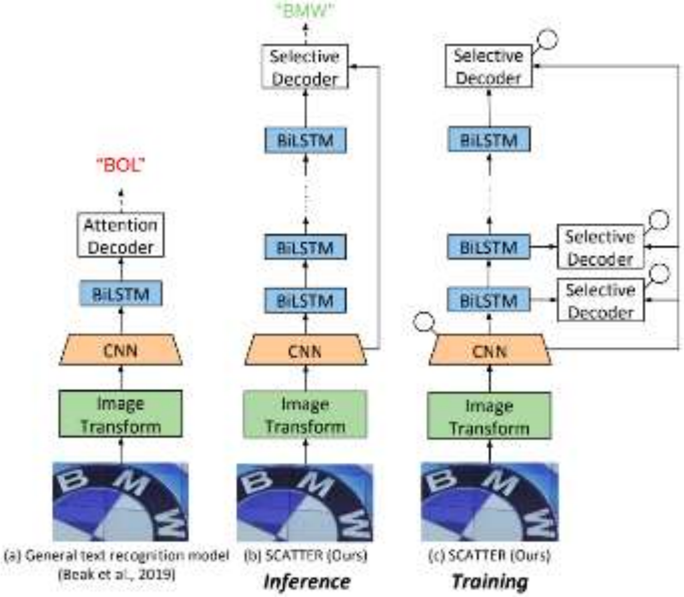
Single-scale testing	Det.	End-to-end			Word spotting		
	F	S	W	G	S	W	G
HUST MCLAB [33][34]	75	67.86	*	*	70.57	*	*
Deep TextSpotter [ICCV'17]	*	54.0	51.0	47.0	58.0	53.0	51.0
Mask TextSpotter [ECCV'18]	86	79.3	73.0	**62.4**	79.3	74.5	**64.2**
FOTS [CVPR'18]	**87.99**	81.09	75.9	60.8	84.64	79.32	63.29
End2End	87.01	81.08	75.09	59.93	84.62	78.35	60.92
End2End-PSL	87.24	**81.18**	**76.25**	62.29	**84.77**	**79.74**	63.91

- OCR #2

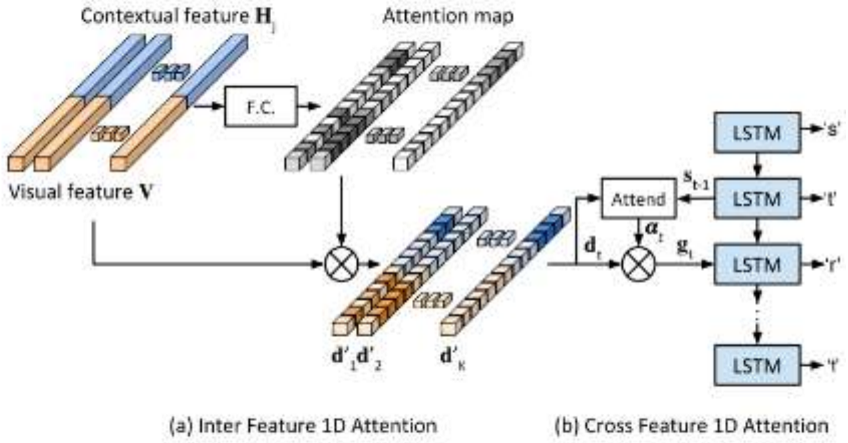
SCATTER: Selective Context Attentional Scene Text Recognizer

- Abstract & Architecrures

SCATTER training and inference architecture



Two-Step Attention Selective Decoder



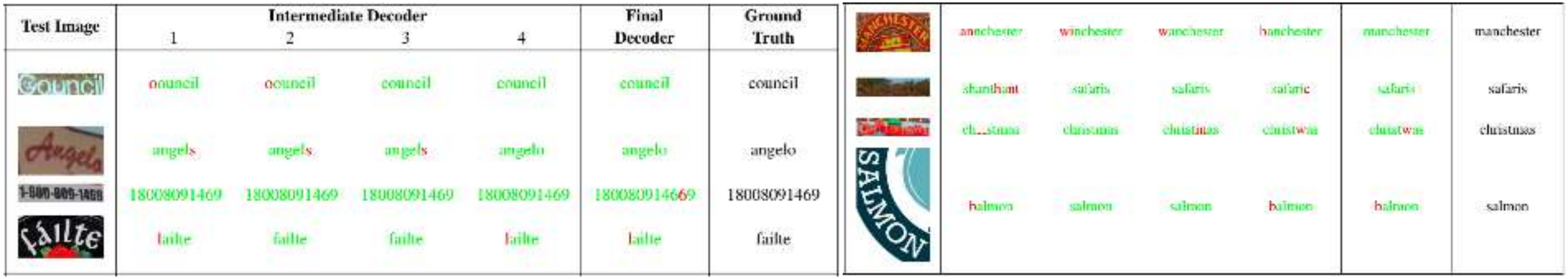
Scene Text Recognition (STR), the task of recognizing text against complex image backgrounds, is an active area of research. Current state-of-the-art (SOTA) methods still struggle to recognize text written in arbitrary shapes. In this paper, we introduce a novel architecture for STR, named **Selective Context ATtentional Text Recognizer (SCATTER)**. SCATTER utilizes a stacked block architecture with intermediate supervision during training, that paves the way to successfully train a deep BiLSTM encoder, thus improving the encoding of contextual dependencies. Decoding is done using a two-step 1D attention mechanism. The first attention step re-weights visual features from a CNN backbone together with contextual features computed by a BiLSTM layer. The second attention step, similar to previous papers, treats the features as a sequence and attends to the intra-sequence relationships.

CVPR 2020. Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R. Manmatha.

- OCR #2

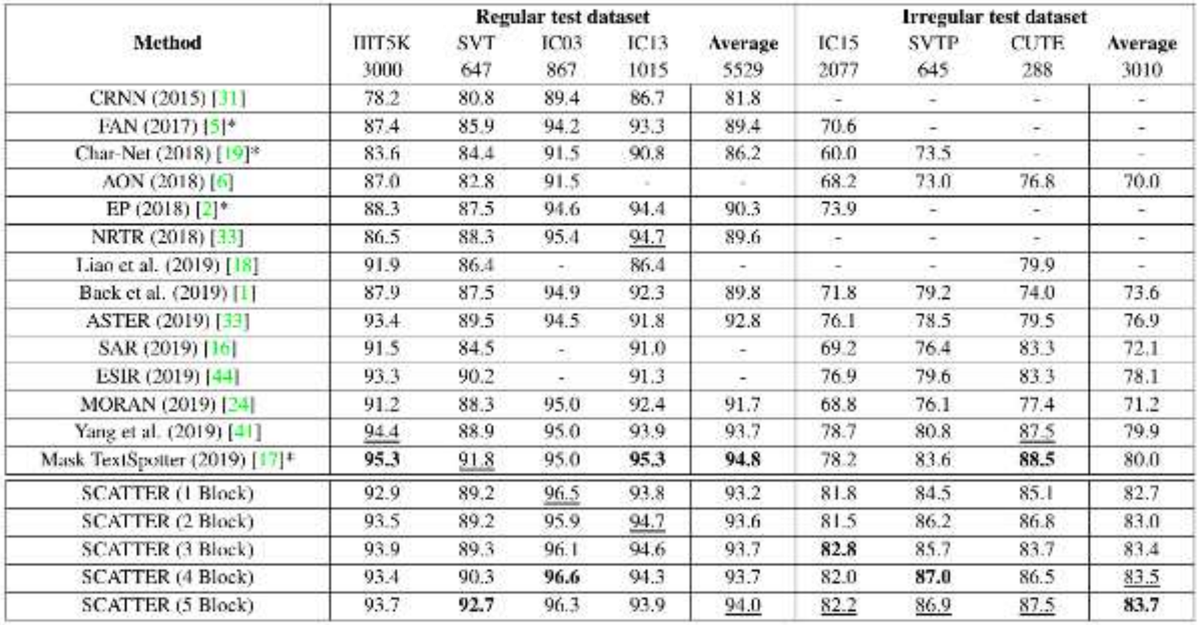
SCATTER: Selective Context Attentional Scene Text Recognizer

- Results



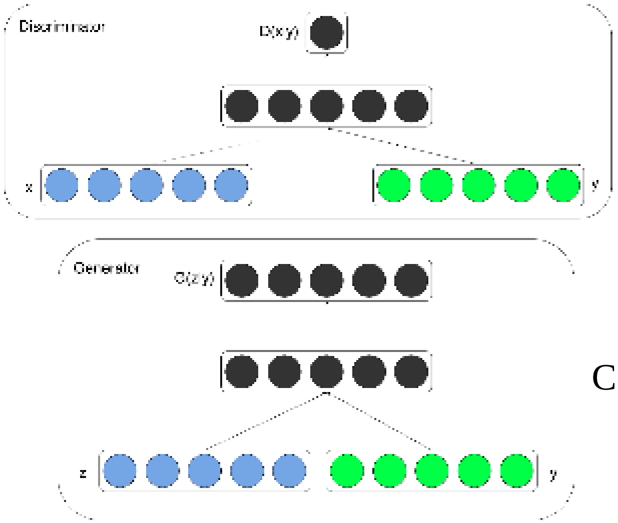
Test Image	Intermediate Decoder 1	2	3	4	Final Decoder	Ground Truth
	oouncil	oouncil	council	council	council	council
	angels	angels	angels	angelo	angelo	angelo
	18008091469	18008091469	18008091469	18008091469	180080914669	18008091469
	lailte	failte	failte	lailte	lailte	failte
	annchester	winchester	wanchester	hanchester	manchester	manchester
	shanthant	safaris	safaris	safaric	safaris	safaris
	ch__stmas	christmas	christmas	christwas	christwas	christmas
	halmon	salmon	salmon	halmon	halmon	salmon

Table 1: Scene text recognition accuracies (%) over seven public benchmark test datasets (number of words in each dataset are shown below the title). No lexicon is used. In each column, the best performing result is shown in **bold** font, and the second best result is shown with an underline. Average columns are weighted (by size) average results on the regular and irregular datasets. "*" indicates using both word-level and character-level annotations for training.



Method	Regular test dataset					Irregular test dataset			
	IIIT5K 3000	SVT 647	IC03 867	IC13 1015	**Average 5529**	IC15 2077	SVTP 645	CUTE 288	**Average 3010**
CRNN (2015) [31]	78.2	80.8	89.4	86.7	81.8	-	-	-	-
FAN (2017) [5]*	87.4	85.9	94.2	93.3	89.4	70.6	-	-	-
Char-Net (2018) [19]*	83.6	84.4	91.5	90.8	86.2	60.0	73.5	-	-
AON (2018) [6]	87.0	82.8	91.5	-	-	68.2	73.0	76.8	70.0
EP (2018) [2]*	88.3	87.5	94.6	94.4	90.3	73.9	-	-	-
NRTR (2018) [33]	86.5	88.3	95.4	94.7	89.6	-	-	-	-
Liao et al. (2019) [18]	91.9	86.4	-	86.4	-	-	-	79.9	-
Baek et al. (2019) [1]	87.9	87.5	94.9	92.3	89.8	71.8	79.2	74.0	73.6
ASTER (2019) [33]	93.4	89.5	94.5	91.8	92.8	76.1	78.5	79.5	76.9
SAR (2019) [16]	91.5	84.5	-	91.0	-	69.2	76.4	83.3	72.1
ESIR (2019) [44]	93.3	90.2	-	91.3	-	76.9	79.6	83.3	78.1
MORAN (2019) [24]	91.2	88.3	95.0	92.4	91.7	68.8	76.1	77.4	71.2
Yang et al. (2019) [41]	94.4	88.9	95.0	93.9	93.7	78.7	80.8	87.5	79.9
Mask TextSpotter (2019) [17]*	**95.3**	91.8	95.0	**95.3**	**94.8**	78.2	83.6	**88.5**	80.0
SCATTER (1 Block)	92.9	89.2	96.5	93.8	93.2	81.8	84.5	85.1	82.7
SCATTER (2 Block)	93.5	89.2	95.9	94.7	93.6	81.5	86.2	86.8	83.0
SCATTER (3 Block)	93.9	89.3	96.1	94.6	93.7	**82.8**	85.7	83.7	83.4
SCATTER (4 Block)	93.4	90.3	**96.6**	94.3	93.7	82.0	**87.0**	86.5	83.5
SCATTER (5 Block)	93.7	**92.7**	96.3	93.9	94.0	82.2	86.9	87.5	**83.7**

GAN



CGAN

GAN：对抗生成网络

应用领域	半监督学习、图像生成/转换、隐空间...
代表算法	1.Ian J. Goodfellow, Mehdi Mirza(蒙特利尔)——GAN、SAGAN、CGAN
	2.Dacheng Tao(USYD)——PAN、TDGAN、GcGAN、E-GAN
	3.Alexei A. Efros(UC Berkeley)——CycleGAN、Pix2pix
	4.NVIDIA——GauGAN
	5.Jiwon Kim(SK T-Brain)——DiscoGAN
	6.Minglun Gong(纽芬兰纪念大学)——DualGAN
	7.Bo Zhang(清华大学)——Triple-GAN

- Generative Adversarial Networks (GAN)

近年来，生成对抗网络（GAN）是一个热门的研究课题。2014 年至今，人们对 GAN 进行了广泛的研究，并提出了大量算法。但是，很少有全面的研究来解释不同 GAN 变体之间的联系以及它们演变的方式。在本文中，我们尝试从算法、理论和应用的角度对多种 GAN 方法进行综述。首先，我们详细介绍了大多数 GAN 算法的研究动机、数学表征和架构。此外，GAN 已经在一些特定应用上与其它机器学习算法相结合，如半监督学习、迁移学习和强化学习。本文比较了这些 GAN 方法的异同。其次，我们研究了与 GAN 相关的理论问题。第三，我们阐述了 GAN 在图像处理与计算机视觉、自然语言处理、音乐、语音与音频、医学以及数据科学中的典型应用。最后，我们指出了 GAN 的一些未来的开放性研究问题。

arXiv 2020. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications

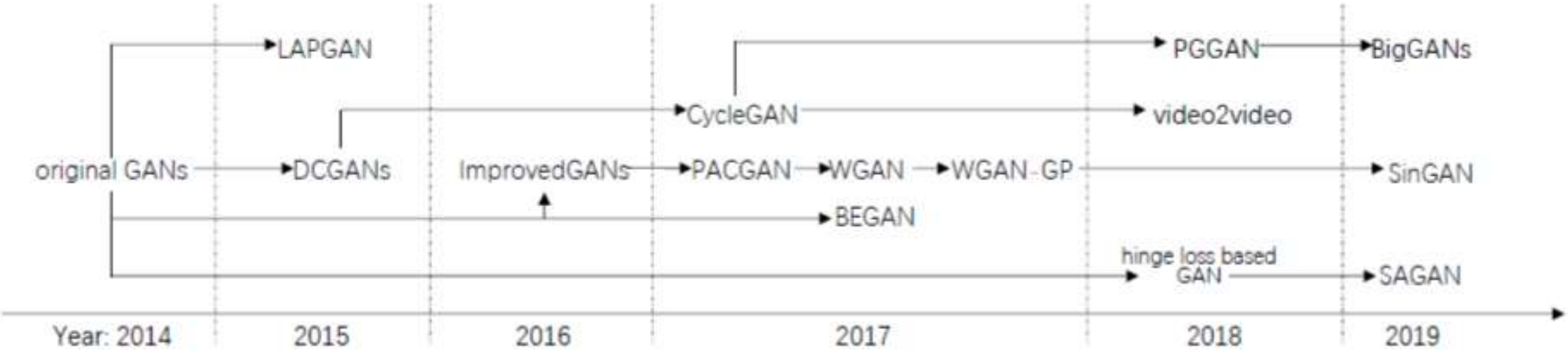


Fig. 5: A road map of GANs. Milestone variants are shown in this figure.

https://github.com/hindupuravinash/the-gan-zoo

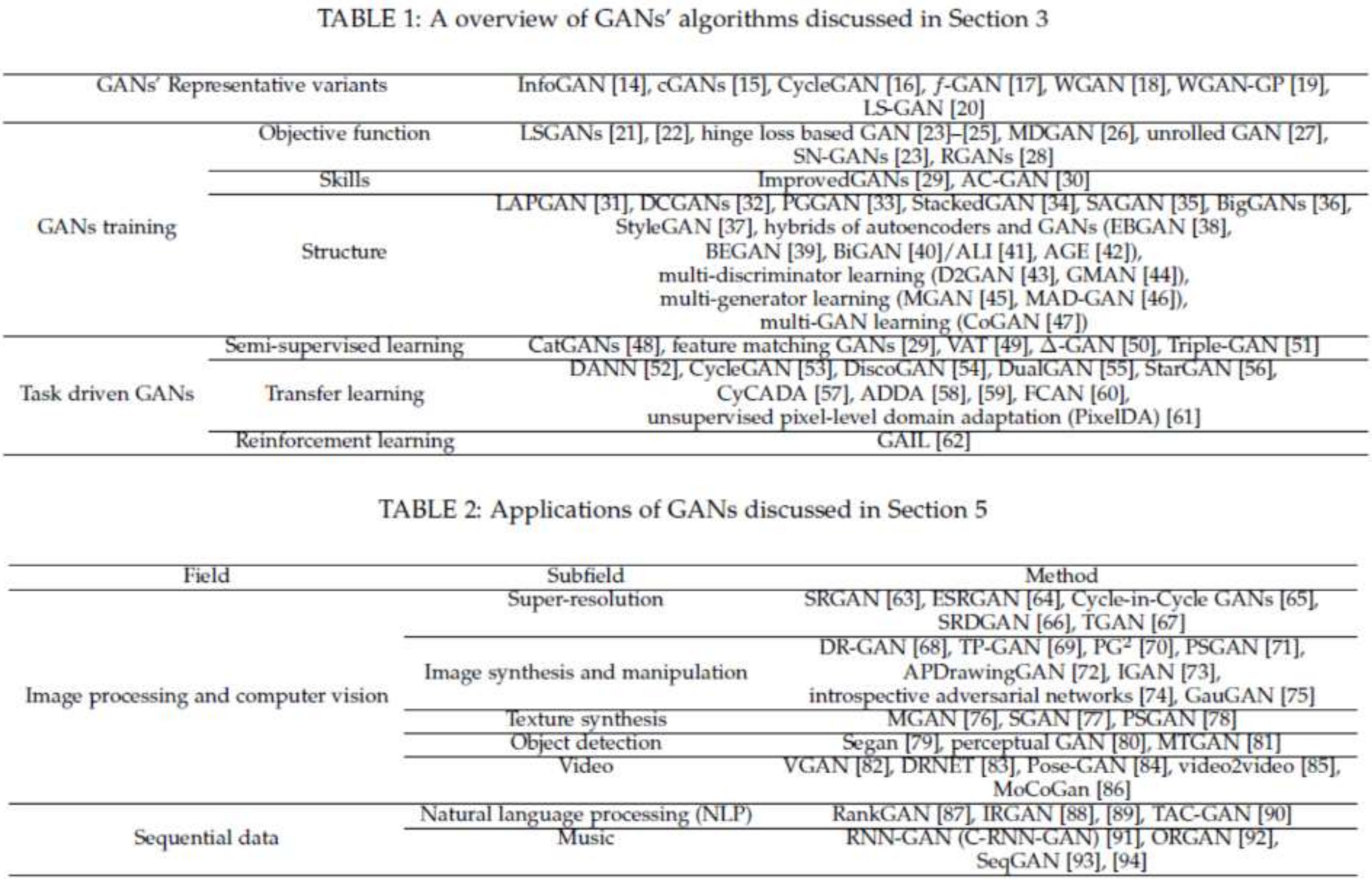


TABLE 1: A overview of GANs' algorithms discussed in Section 3

GANs' Representative variants		InfoGAN [14], cGANs [15], CycleGAN [16], f-GAN [17], WGAN [18], WGAN-GP [19], LS-GAN [20]
GANs training	Objective function	LSGANs [21], [22], hinge loss based GAN [23]–[25], MDGAN [26], unrolled GAN [27], SN-GANs [23], RGANs [28]
	Skills	ImprovedGANs [29], AC-GAN [30]
	Structure	LAPGAN [31], DCGANs [32], PGGAN [33], StackedGAN [34], SAGAN [35], BigGANs [36], StyleGAN [37], hybrids of autoencoders and GANs (EBGAN [38], BEGAN [39], BiGAN [40]/ALI [41], AGE [42]), multi-discriminator learning (D2GAN [43], GMAN [44]), multi-generator learning (MGAN [45], MAD-GAN [46]), multi-GAN learning (CoGAN [47])
Task driven GANs	Semi-supervised learning	CatGANs [48], feature matching GANs [29], VAT [49], Δ-GAN [50], Triple-GAN [51]
	Transfer learning	DANN [52], CycleGAN [53], DiscoGAN [54], DualGAN [55], StarGAN [56], CyCADA [57], ADDA [58], [59], FCAN [60], unsupervised pixel-level domain adaptation (PixelDA) [61]
	Reinforcement learning	GAIL [62]

TABLE 2: Applications of GANs discussed in Section 5

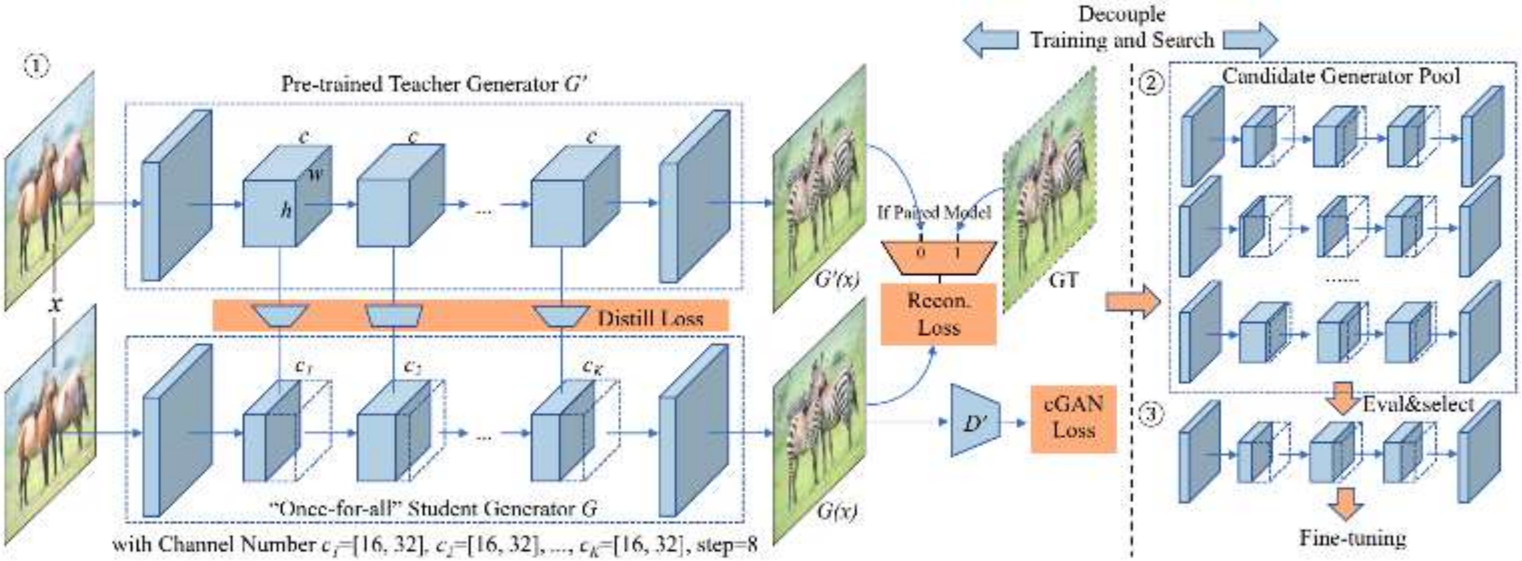
Field	Subfield	Method
Image processing and computer vision	Super-resolution	SRGAN [63], ESRGAN [64], Cycle-in-Cycle GANs [65], SRDGAN [66], TGAN [67]
	Image synthesis and manipulation	DR-GAN [68], TP-GAN [69], PG^2 [70], PSGAN [71], APDrawingGAN [72], IGAN [73], introspective adversarial networks [74], GauGAN [75]
	Texture synthesis	MGAN [76], SGAN [77], PSGAN [78]
	Object detection	Segan [79], perceptual GAN [80], MTGAN [81]
	Video	VGAN [82], DRNET [83], Pose-GAN [84], video2video [85], MoCoGan [86]
Sequential data	Natural language processing (NLP)	RankGAN [87], IRGAN [88], [89], TAC-GAN [90]
	Music	RNN-GAN (C-RNN-GAN) [91], ORGAN [92], SeqGAN [93], [94]

arXiv 2020. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications

- GAN #1

GAN Compression: Efficient Architectures for Interactive Conditional GANs

- Abstract & Framework



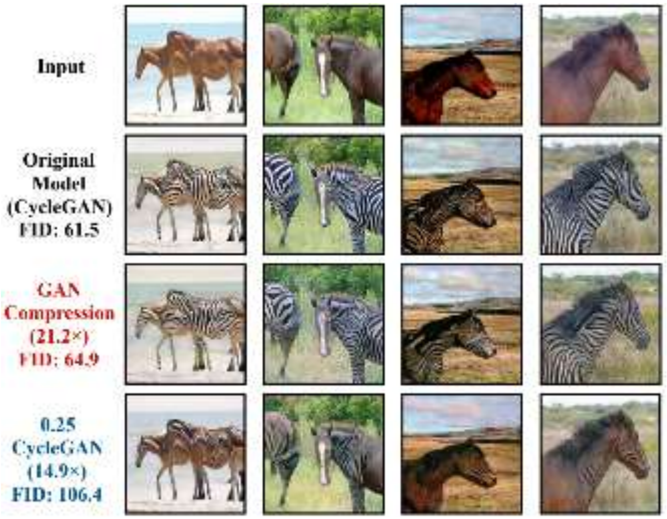
Conditional Generative Adversarial Networks (cGANs) have enabled controllable image synthesis for many computer vision and graphics applications. However, recent cGANs are 1-2 orders of magnitude more computationally-intensive than modern recognition CNNs. For example, GauGAN consumes 281G MACs per image, compared to 0.44G MACs for MobileNet-v3, making it difficult for interactive deployment. In this work, we propose **a general-purpose compression framework** for reducing the inference time and model size of the generator in cGANs. Directly applying existing CNNs compression methods yields poor performance due to the difficulty of GAN training and the differences in generator architectures. We address these challenges in two ways. First, to stabilize the GAN training, we transfer knowledge of multiple intermediate representations of the original model to its compressed model, and unify unpaired and paired learning. Second, instead of reusing existing CNN designs, our method automatically finds efficient architectures via neural architecture search (NAS). To accelerate the search process, we decouple the model training and architecture search via weight sharing

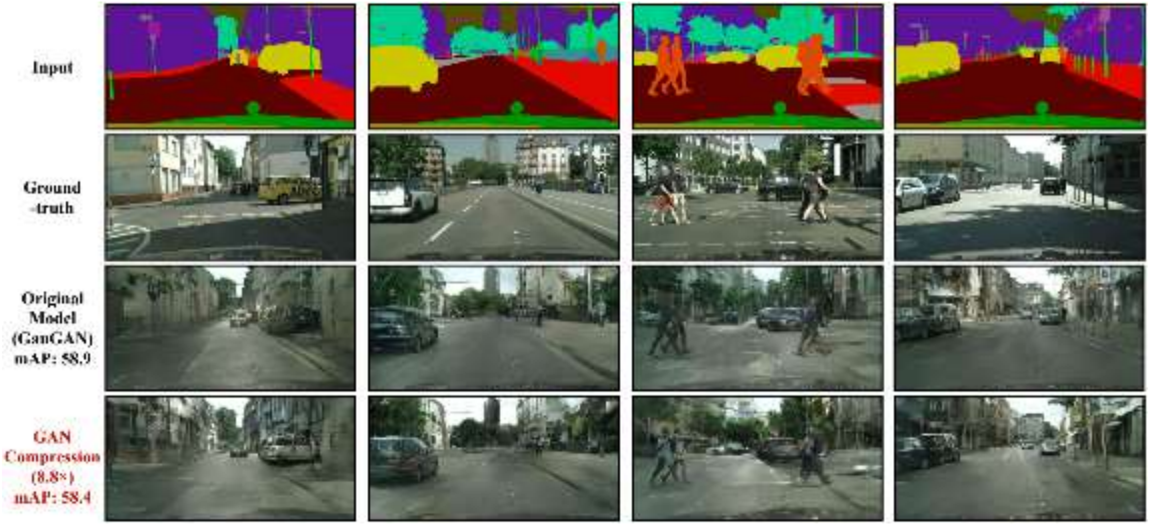
CVPR 2020. Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han.

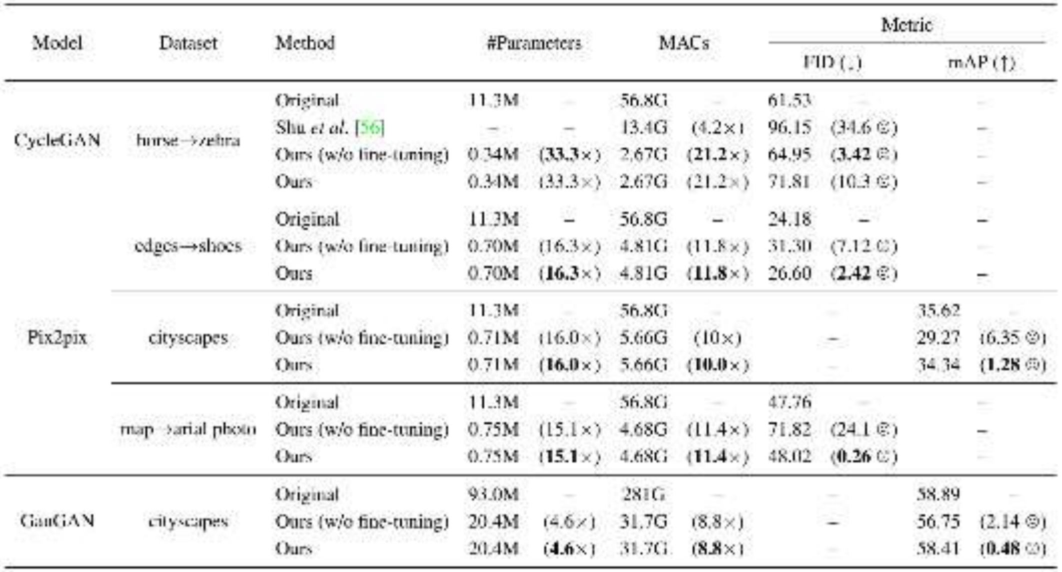
- GAN #1

GAN Compression: Efficient Architectures for Interactive Conditional GANs

- Resutls


Input
Original Model (CycleGAN) FID: 61.5
GAN Compression (21.2×) FID: 64.9
0.25 CycleGAN (14.9×) FID: 106.4


Input
Ground -truth
Original Model (GauGAN) mAP: 58.9
GAN Compression (8.8×) mAP: 58.4

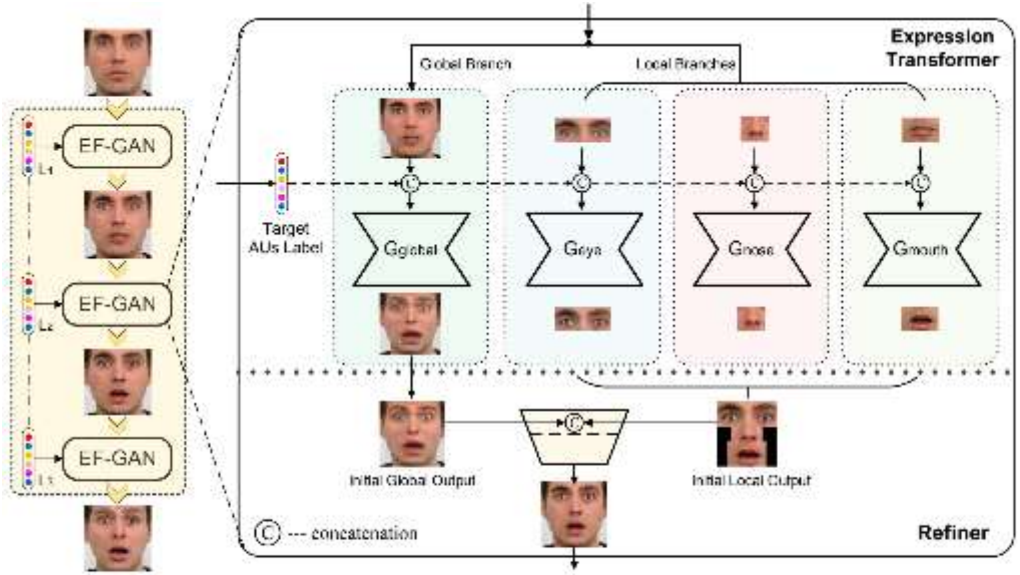


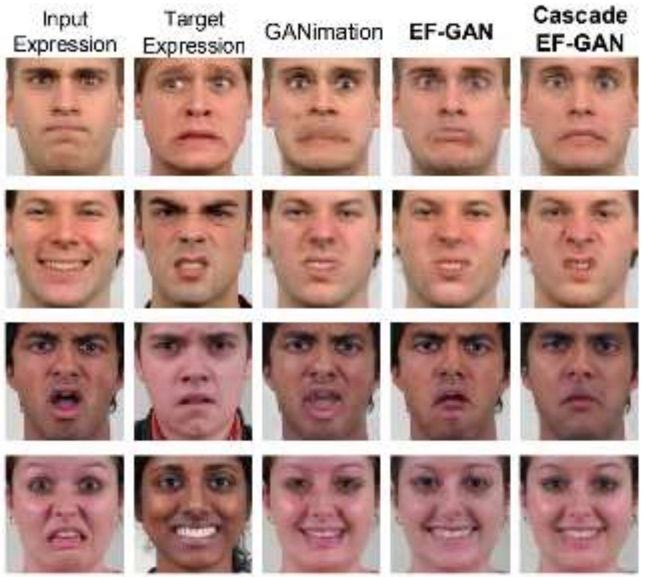
Model	Dataset	Method	#Parameters		MACs		Metric			
							FID (↓)		mAP (↑)	
CycleGAN	horse→zebra	Original	11.3M	–	56.8G	–	61.53	–	–	
		Shu *et al.* [56]	–	–	13.4G	(4.2×)	96.15	(34.6 ☹)	–	
		Ours (w/o fine-tuning)	0.34M	**(33.3×)**	2.67G	**(21.2×)**	64.95	(**3.42** ☹)	–	
		Ours	0.34M	(33.3×)	2.67G	(21.2×)	71.81	(10.3 ☹)	–	
	edges→shoes	Original	11.3M	–	56.8G	–	24.18	–	–	
		Ours (w/o fine-tuning)	0.70M	(16.3×)	4.81G	(11.8×)	31.30	(7.12 ☹)	–	
		Ours	0.70M	(**16.3**×)	4.81G	(**11.8**×)	26.60	(**2.42** ☹)	–	
Pix2pix	cityscapes	Original	11.3M	–	56.8G	–	–		35.62	–
		Ours (w/o fine-tuning)	0.71M	(16.0×)	5.66G	(10×)	–		29.27	(6.35 ☹)
		Ours	0.71M	(**16.0**×)	5.66G	(**10.0**×)	–		34.34	(**1.28** ☹)
	map→arial photo	Original	11.3M	–	56.8G	–	47.76	–	–	
		Ours (w/o fine-tuning)	0.75M	(15.1×)	4.68G	(11.4×)	71.82	(24.1 ☹)	–	
		Ours	0.75M	(**15.1**×)	4.68G	(**11.4**×)	48.02	(**0.26** ☹)	–	
GauGAN	cityscapes	Original	93.0M	–	281G	–	–		58.89	–
		Ours (w/o fine-tuning)	20.4M	(4.6×)	31.7G	(8.8×)	–		56.75	(2.14 ☹)
		Ours	20.4M	(**4.6**×)	31.7G	(**8.8**×)	–		58.41	(**0.48** ☹)

- GAN #2

Cascade EF-GAN: Progressive Facial Expression Editing with Local Focuses

- Abstract, Architecture & Results





Recent advances in Generative Adversarial Nets (GANs) have shown remarkable improvements for facial expression editing. However, current methods are still prone to generate artifacts and blurs around expression-intensive regions, and often introduce undesired overlapping artifacts while handling large-gap expression transformations such as transformation from furious to laughing. To address these limitations, we propose **Cascade Expression Focal GAN (Cascade EF-GAN)**, a novel network that performs progressive facial expression editing with local expression focuses. The introduction of the local focus enables the Cascade EF-GAN to better preserve identity-related features and details around eyes, noses and mouths, which further helps reduce artifacts and blurs within the generated facial images. In addition, an innovative cascade transformation strategy is designed by dividing a large facial expression transformation into multiple small ones in cascade, which helps suppress overlapping artifacts and produce more realistic editing while dealing with large-gap expression transformations.

CVPR 2020. Rongliang Wu, Gongjie Zhang, Shijian Lu*, and Tao Chen.

Thanks for Listening!

Dr. Lefei Zhang

zhanglefei@whu.edu.cn